UNIVERSITY OF CALIFORNIA

Los Angeles

SINGLE AND DUAL-ENERGY COMPUTED TOMOGRAPHY GUIDED RADIATION THERAPY

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Physics and Biology in Medicine

by

Qihui Lyu

2021

ABSTRACT OF THE DISSERTATION

SINGLE AND DUAL-ENERGY COMPUTED TOMOGRAPHY GUIDED RADIATION THERAPY

by

Qihui Lyu

Doctor of Philosophy in Physics and Biology in Medicine

University of California, Los Angeles, 2021

Professor Ke Sheng, Committee Chair

Purpose

Conformal dose and precise imaging are key to radiation therapy. Here we introduce a series of integrated optimization frameworks to improve computed tomography (CT) image quality, refine dual-energy CT (DECT) material decomposition, and advance treatment planning and delivery methods.

Methods

We formulate our optimization framework as a least-square fidelity term and a regularization term. The regularization term was designed specifically for each application, including accounting for the piecewise smoothness of the CT image, sparsity in the DECT decomposition image, and mechanical constraints in radiotherapy. The flexible optimization framework allows novel treatments that unleash unnecessary constraints in the current Volumetric Modulated Arc Therapy (VMAT) or Intensity Modulated Radiation Therapy (IMRT), including allowing dual-layer Multi-Layer Collimator (DLMLC), non-isocentric treatment, dynamic collimator, non-coplanar beams, and FLASH radiotherapy. The added degrees of freedom significantly expand the searching space, and these large-scale optimization problems were solved with a Fast Iterative Shrinkage-Thresholding Algorithm (FISTA).

Results

On a Catphan study, our integrated CT reconstruction distinguishes a higher number of line pairs and preserves low contrast objects compared with conventional Filtered Back Projection (FBP) and total variation (TV) reconstruction. Our framework improved the decomposition accuracy from 63.9% to 99.8% when applying for material decomposition compared with the classic direct inversion method. Compared with single-layer MLC (SLMLC) VMAT, DLMLC VMAT reduced R50 by 10% for radiotherapy treatment. Compared with isocentric 100cm-source-to-isocenter distance (SID-100cm) 4πIMRT, the non-isocentric SID-50cm 4πIMRT reduced R50 and integral dose by 5.3% and 9.6%. Compared with static collimator VMAT (SCVMAT), dynamic collimator VMAT (DCVMAT) reduced the max and mean organs-at-risk (OAR) dose by 4.49% and 2.53% of the prescription dose. Compared with coplanar VMAT, 4πVMAT reduced R50 by 19.7%. Compared with clinical VMAT, our FLASH delivery method reduced the max and mean OAR physical doses by 4.8Gy and 6.3Gy in addition to potential biological gains.

Conclusions

The integrated optimization framework improves CT image quality, DECT decomposition accuracy, and radiotherapy dose conformality.

The dissertation of Qihui Lyu is approved.

Dan Ruan

Michael McNitt-Gray

Steve Jiang

Wotao Yin

Ke Sheng, Committee Chair

University of California, Los Angeles

2021

To Fangming, Janna, and Mom

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

ACKNOWLEDGEMENTS

First and foremost, my greatest thank goes to my advisor Dr. Ke Sheng. He is the best mentor I can ever ask for. He always has very creative ideas and insightful views about the field, while he is also an uncommonly brilliant educator. He lets us lead our own projects, encourages us to explore and learn, and enlightens us to become independent researchers. Meanwhile, he is extremely kind, considerate, understanding, and supportive. He always trusts me even when I am not confident about myself. I feel extremely lucky and privileged to work with him.

I would also like to thank Dr. Dan Ruan for her constant guidance and support throughout my graduate career. She has contributed to many core ideas in this thesis. Talking with her is always fun and inspiring.

I would also like to thank Dr. Steve Jiang and Dr. Wotao Yin on my dissertation committee. I benefited a lot from discussing my projects with them. Dr. Yin's class on monotone operators helps me understand optimization from a theoretical perspective.

I would also like to thank Dr. Michael McNitt-Gray, for being a dissertation committee member, a valuable collaborator, and a dedicated program director. The UCLA PBM is such a unique and wonderful graduate program. I would also like to thank Reth and Alondra for everything they did for our program.

Thanks should also go to my other collaborators, Dr. Jessica Miller, Dr. Arion Chatziioannou, and Dr. Tianye Niu, for their precious data and valuable insight.

I could not have done this without my wonderful labmates, Dan, Victoria, Angelia, Kaley, Wenbo, Ryan, Elizabeth, Daili, Jiayi, Pav, Qifan, Lu, Alan, Flute, Holly, Daniel, Yi,

Ningning, Yang, and Nuo. It has been a great fortune to be a part of the team and witness many exciting research and discoveries. Beyond that, their friendship and companion made the journey full of joy. In particular, I want to thank Dan who introduced me to optimization and medical physics. Thank Daniel for being the optimization expert. Thank Ryan for his support on many projects in this thesis. Thank Wenbo and Victoria for being both my best colleagues and best friends.

Thanks should also go to my classmates Jie, Bao, Elizabeth, Nyasha, Alex, Vahid, Rachel, and my friends Zhuo and Wenxi. Their friendship and companion make me feel settled down when I first came here from abroad to start graduate school.

I would also like to thank my mom Yanmei and dad Wenliang for their unconditional love and support. They let me wander on my own and trust all my decisions. Thank you to my daughter Janna, who was born three months ago. Her happy face lights up my world.

Finally, I want to thank my husband Fangming. He is not only my partner, but also my best friend, a person who knows me the most, with whom I can share absolutely everything. Thank you for supporting all my crazy ideas. You made me who I am today.

VITA

EDUCATION

M.S.	University of California Los Angeles, Physics and Biology in Medicine	2019
B.S.	Nanjing University, Physics	2016

AWARDS

AAPM Science Council Associates Mentorship Program	2021
Dr. Moses A. Greenfield Award	2020
Dissertation Year Fellowship	2020
AAPM Expanding Horizons Travel Grant	2019
AAPM Science Council - Selected Speaker	2018
AAPM Science Highlights - Selected Speaker	2018

JOURNAL PUBLICATIONS

1. **Lyu, Q.,** Neph, R., and Sheng, K. (2021), Pair-Production Tomography Imaging, in revision
2. Neph, R., **Lyu, Q.,** Yang, Y., and Sheng, K. (2021), DeepMC: A Deep Learning Method for Efficient Monte Carlo Beamlet Dose Calculation by Predictive Denoising in Magnetic Resonance-guided Radiotherapy, Phys. Med. Biol. 66 035022 doi: 10.1088/1361-6560/abca01
3. **Lyu, Q.,** Neph, R., Ruan, D., Boucher, S., and Sheng, K. (2021), ROAD: ROtational direct Aperture optimization with a Decoupled ring-collimator for FLASH radiotherapy, Phys. Med. Biol. 66 035020 doi: 10.1088/1361-6560/abcbd0
4. Gu, W., Ruan, D., **Lyu, Q.,** Zou, W., Dong, L., and Sheng, K. (2020), A Novel Energy Layer Optimization Framework with for Spot-Scanning Proton Arc Therapy, Med. Phys., 47: 2072-2084. doi:10.1002/mp.14083
5. Jiang Y, Zhang X, Sheng K, Niu T, Xue Y, **Lyu Q,** et al. (2020), Noise Suppression in Image-Domain Multi-Material Decomposition for Dual-Energy CT. IEEE Trans Biomed Eng. 2020;67: 523–535. 10.1109/TBME.2019.2916907
6. **Lyu, Q.,** Neph, R., Yu, V., Ruan, D., Boucher, S., and Sheng, K. (2020), Many-isocenter Optimization for Robotic Radiotherapy, Phys. Med. Biol. Volume 65, Number 4 pp:5003 doi: 10.1088/1361-6560/ab63b8
7. **Lyu, Q.,** O'Connor, D., Niu, T. and Sheng, K. (2019), Image-domain multi-material decomposition for dual-energy CT with non-convex sparsity regularization, J. Med. Imag. 6(4), 044004, doi: 10.1117/1.JMI.6.4.044004
8. **Lyu, Q.,** Neph, R., Yu, V., Ruan, D., and Sheng, K. (2019), Single-arc VMAT optimization for dual-layer MLC, Phys. Med. Biol. vol. 64 issue 9 pp: 095028 doi:10.1088/1361-6560/ab0ddd

9. Xue Y, Jiang Y, Yang C, **Lyu Q,** Wang J, Luo C, et al. (2019), Accurate Multi-Material Decomposition in Dual-Energy CT: A Phantom Study. IEEE Transactions on Computational Imaging. 2019;5: 515–529. doi: 10.1109/TCI.2019.2909192
10. **Lyu, Q.,** Yu, V., Ruan, D., Neph, R., O'Connor, D., and Sheng, K. (2018), A novel optimization framework for VMAT with dynamic gantry couch rotation, Phys. Med. Biol. vol. 63 issue 12 pp: 125013 doi:10.1088/1361-6560/aac704
11. **Lyu, Q.,** O'Connor, D. , Ruan, D. , Yu, V. , Nguyen, D. and Sheng, K. (2018), VMAT optimization with dynamic collimator rotation. Med. Phys., 45: 2399-2410. doi:10.1002/mp.12915
12. **Lyu, Q.,** Yang, C. , Gao, H. , Xue, Y. , O'Connor, D. , Niu, T. and Sheng, K. (2018), Technical Note: Iterative megavoltage CT (MVCT) reconstruction using block-matching 3D-transform (BM3D) regularization. Med. Phys., 45: 2603-2610. doi:10.1002/mp.12916
13. Yang, C. , Wu, P. , Gong, S. ,Wang, J., **Lyu, Q. ,** Tang, X., Niu, T. (2017), Shading correction assisted iterative cone-beam CT reconstruction. Phys. Med. Biol. vol. 62 issue 22 pp: 8495-8520 doi: 10.1088/1361-6560/aa8e62
14. Nguyen, D. , **Lyu, Q.,** Ruan, D. , O'Connor, D. , Low, D. A. and Sheng, K. (2016), A comprehensive formulation for volumetric modulated arc therapy planning. Med. Phys., 43: 4263-4272. doi:10.1118/1.4953832

Conference Proceeding Papers:

1. **Lyu, Q.,** Niu, T., Ruan, D. and Sheng, K. , Iterative reconstruction of cone-beam breast CT using plug-and-play projected gradient descent, Proceedings Volume 11312, Medical Imaging 2020: Physics of Medical Imaging; 113124J (2020) doi:10.1117/12.2549787
2. **Lyu, Q.,** Ruan, D., Hoffman, J., Neph R., McNitt-Gray, M. and Sheng, K. (2019), Iterative reconstruction for low dose CT using Plug-and-Play alternating direction method of multipliers (ADMM) framework, Proc. SPIE 10949, Medical Imaging 2019: Image Processing, 1094906 doi:10.1117/12.2512484
3. **Lyu, Q.,** O'Connor, D., Niu, T. and Sheng, K., Image-domain multi-material decomposition for dual-energy CT with non-convex sparsity regularization, Proc. SPIE 10949, Medical Imaging 2019: Image Processing, 1094903, doi:10.1117/12.2508037

SELECTED CONFERENCE PRESENTATIONS

1. **Q Lyu*,** R Neph, D O'Connor, D Ruan, S Boucher, K Sheng, ROAD: ROtational Direct Aperture Optimization with a Decoupled Ring-Collimator for FLASH, Oral presentation at AAPM 2020
2. **Q Lyu*,** D Ruan , J Hoffman , R Neph , M McNitt-Gray , K Sheng, , Iterative reconstruction for low dose CT using Plug-and-Play ADMM framework, Oral presentation at SPIE medical imaging 2019
3. **Q Lyu*,** V Yu, D O'Connor, R Neph, D Ruan, K Sheng. 4πVMAT for Efficient Delivery of Highly Conformal Non-Coplanar Plans, Oral presentation at AAPM 2018
4. **Q Lyu*,** D O'Connor, T Niu, K Sheng. A Novel Multi-Material Decomposition Method for Dual Energy CT, Oral presentation at AAPM 2018

1 INTRODUCTION

1.1 Radiation therapy

Cancer is a leading cause of death worldwide, accounting for an estimated 9.6 million deaths each year[1]. 60% of cancer patients received radiotherapy treatment[2]. Radiotherapy, also known as radiation therapy, uses high-energy radiation beam or radioactive substances to induce DNA damage to cells and subsequently inhibit the growth, division, metastasis, and proliferation of malignant tumor cells. The ionizing radiation used in radiotherapy treatment is a “double-edged sword”, since it not only affects the biological processes of neoplasms but also injures normal tissue cells, leading to undesired treatment-related side effects.

To maximize the effectiveness of radiotherapy treatment, it is essential to accurately deliver a conformal dose distribution to the patient, where the tumor receives a prescription dose and organs-at-risk (OAR) dose is minimized. The high accuracy and high conformality can only be achieved with proper image guidance and optimized treatment planning, demanding optimality in each of the four steps of radiotherapy treatment: imaging, tumor and OAR delineation, treatment planning, and image-guided treatment delivery. This thesis develops a suite of novel treatment techniques and optimization algorithms that address challenges arising from multiple stages of radiotherapy (Figure 1-1). The thesis is structured into three parts. Chapter 2 proposes an integrated Computed Tomography (CT) iterative reconstruction method that improves both diagnostic CT images and the image guidance during treatment delivery. The method was applied to lung CT images for lung cancer screening, Cone-beam breast CT (CBBCT) images for breast cancer screening, and Megavoltage CT (MVCT) for onboard imaging verification. Chapter 3 describes dual-energy CT (DECT) multi-material decomposition method to enhance tissue characterization and differentiation. The method was applied to differentiate the radiosensitive active bone marrow for dose sparing. Chapter 4 entails a flexible optimization framework for a variety of novel treatment planning and delivery techniques.

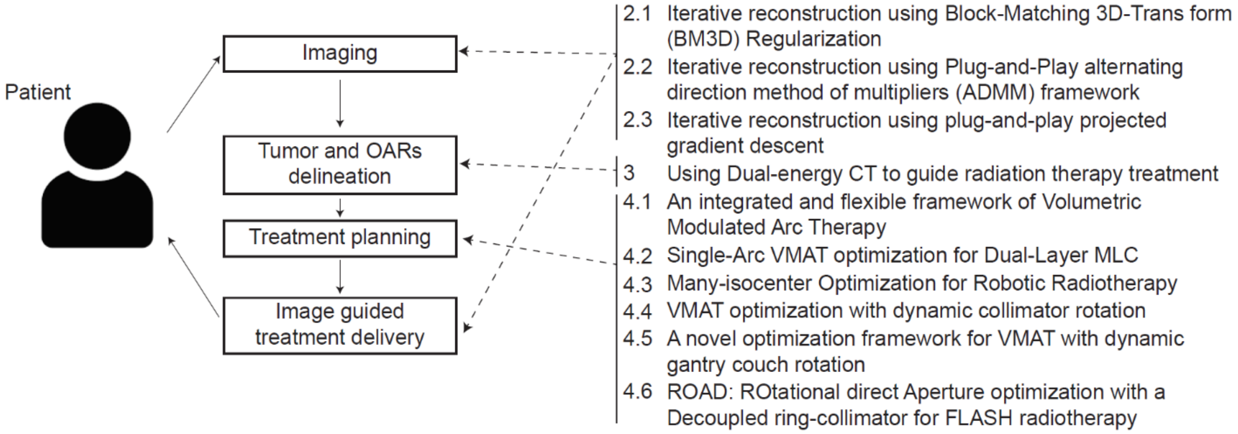


Figure 1-1 A typical workflow of radiotherapy and the corresponding contributions from this thesis.

1.2 Computed Tomography

Compared with other medical imaging techniques such as magnetic resonance imaging (MRI)[3–5] and Positron emission tomography (PET)[6], CT[7,8] using X-rays imaging systems is advantageous in its low cost, high stability, high speed, high resolution, and high sensitivity to dense or high-atomic number materials. CT is indispensable in many radiotherapy applications, including cancer diagnosis, structure delineation, treatment planning dose evaluation, image guidance during treatment[9].

Despite the wide adoption of CT, its ionizing radiation can cause potential detrimental effects, including cancer and genetic disease. This has raised growing concern on using diagnostic CT scan for large-scale population screening and on repeated CT scans in image-guided radiation therapy, which motivated the development of low dose CT (ldCT). In practice, imaging dose reduction is achieved using sparse projection sampling and/or reduced x-ray tube current[10,11].

A 3D CT image can be reconstructed from 2D projection images from many different angles around the patient using the analytical Filtered Back Projection (FBP) algorithm or iterative reconstruction algorithm[12,13]. One dilemma in CT reconstruction is the incompatibility of the iterative reconstruction framework and state-of-the-art image denoisers. On the one hand, it can be extremely difficult, if not impossible, to formulate an arbitrary image denoiser as an optimization problem for iterative reconstruction. On the other hand, the ldCT FBP images contain severe image artifacts and noise that low-contrast objects and fine structures cannot be recovered. In Chapter 2, we propose a flexible framework that integrates state-of-the-art image denoisers into CT iterative reconstruction.

1.3 Dual-energy CT (DECT)

Conventional x-ray CT projections are acquired with a single-energy spectrum. The reconstructed single-energy CT (SECT) provides the linear attenuation coefficients (LAC) of the imaging object, but the LACs depend on both the effective atomic number and the electron density, making it insufficient to determine material components. DECT is acquired with two distinct energy spectra that are attenuated differently by the tissues. Therefore, DECT provides enhanced information to better differentiate and quantify material compositions. DECT shows promise in many clinical applications, including virtual unenhancement (VUE) imaging[14,15], liver lesion characterization[16], kidney stone characterization[17], oncologic imaging[14,18], bone removal[14], etc.

Despite the potential of using DECT for multi-material decomposition (MMD), the problem is ill-defined for decomposing more than two materials without additional assumption[19,20]. The classic direct inversion (DI) method[21] cannot accurately separate more than two basis materials and significantly amplifies image noise. Chapter 3 shows a novel integrated MMD method that addresses the piecewise smoothness and intrinsic sparsity property of the decomposition. We also apply it to separate the radiosensitive active bone marrow from fatty bone marrow.

1.4 Radiotherapy treatment planning and delivery

After imaging and target delineation, the next goal is to deliver a conformal dose to the target with treatment planning and delivery. Volumetric modulated arc therapy (VMAT)[22,23] is a widely employed radiation therapy treatment technique due to its significantly higher delivery efficiency than the static beam Intensity Modulated Radiation Therapy (IMRT)[24]. In its current implementation, VMAT delivers radiation dose while the gantry moves continuously on the coplanar arc and the Multi-leaf Collimator (MLC) leaves move simultaneously to shape the beam. At each beam, VMAT delivers a uniform intensity to the target, and all beams collectively form a conformal dose distribution.

Treatment planning is solved via inverse optimization. Compared with the static beam IMRT[24], the VMAT optimization problem is more challenging due to the large problem size and complex mechanical constraints. Typically, 180 or more beams are included in the VMAT delivery compared with fewer than ten beams used in a typical IMRT plan. More importantly, the gantry rotation and leaf motion are coupled: the MLC leaf movements between adjacent beams are restricted by the maximal leaf speed for efficient VMAT delivery.

The VMAT optimization in most commercial planning systems is based on the progressive sampling optimization (PSO) method proposed by Otto[22]. PSO produces deliverable VMAT plans that meet the mechanical constraint of existing delivery methods and machine hardware, but the dosimetry is suboptimal, and the algorithm is incompatible with novel delivery techniques and emerging machine hardware. For example, although collimator rotation is an optimization variable that can be exploited for dosimetric advantages, existing VMAT optimization uses a fixed collimator angle in each arc and only rotates the collimator between arcs. Besides, existing VMAT is typically restricted to coplanar arcs and is dosimetrically inferior to IMRT with optimized non-coplanar beams. The PSO method was also designed for VMAT with a single-layer multi-leaf collimator (SLMLC), which has lower modulation resolution than the dual-layer multi-leaf collimator (DLMLC) system.

To overcome these unnecessary constraints, we propose a flexible VMAT optimization framework with global sampling that improves dosimetry and opens the door to new degrees of freedom in VMAT. We will show in Chapter4 that the framework allows DLMLC and dynamic collimator rotation in a single arc, both of which effectively increase the modulation resolution. The framework also includes non-coplanar trajectories that significantly expand the space of dose spillage and FLASH radiotherapy, which exploits the biological effectiveness of ultrafast radiation.

1.5 Mathematical framework and computation algorithms

In addition to the relevance to radiotherapy, the three parts also share synergies in their mathematical formulation as an integrated optimization framework using compressed sensing technique. Classic algorithms often use heuristic and detached approaches, such as denoising on FBP reconstructed CT images, direct matrix inversion with hard ceiling of three basis for DECT material decomposition, and progressive sampling in VMAT optimization. Such methods are effective in satisfying the hard constraints: removing noise, limiting the number of materials, and VMAT deliverability, but they inevitably compromise the data fidelity. We formulate these problems as an integrated optimization framework of a least square fidelity term and a regularization term, accounting for data consistency and prior knowledge/constraints simultaneously. The regularization term was designed specifically for each application as a soft control. For CT reconstruction, the Block-Matching 3D-Transform (BM3D) term or the ‘plug-and-play’ term encourage image piecewise smoothness, patch-wise correlation, or data-driven image features. For DECT MMD, sparse regularization terms limit the number of decomposition components simultaneously present in each pixel. For VMAT optimization, regularization terms encourage deliverability and delivery efficiency. A single segment term encourages the optimized fluence map to be SLMLC-deliverable or DLMLC-deliverable. A group sparsity term limits the number of isocenters and beams for efficient robotic arm delivery. A group sparsity term, in combination with the Dijkstra's algorithm, promotes smoothness of the dynamic collimator trajectories and dynamic gantry couch trajectories for efficient VMAT delivery.

Both CT iterative reconstruction and VMAT optimization are computationally expensive due to the vast problem size, non-smooth objective, and optimization constraints. For VMAT, the computation complexity is further compounded with added degrees of freedom. Classic convex optimization algorithms such as interior point methods[25] typically have quadratic memory complexities and cubic arithmetic complexities. Solving these problems through interior points methods are computationally intractable. Recent convex optimization research has focused on a class of algorithms known as proximal algorithms[26], which are generally applicable and well-suited for these non-smooth, constrained, large-scale and high-dimensional problems. In particular, we use the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)[27], a member of proximal algorithms, to efficiently solve most of the optimization problems in this thesis. On the one hand, FISTA requires only the multiplication with the system matrix and its transpose at each iteration, which substantially reduces the computational costs as compared with other proximal algorithms such as the alternating direction method of multipliers (ADMM)[28]. On the other hand, it achieves a convergence rate of $\mathrm{O}(\frac{1}{k^2})$, a significant improvement over the $\mathrm{O}(\frac{1}{k})$ convergence rate of most other proximal methods such as ADMM and Primal-dual hybrid gradient method (PDHG)[29].

1.6 Overview

Chapter 2 describes methods to integrate state-of-the-art denoisers in CT iterative reconstruction (IR). Section 2.1 demonstrates a method that formulates the Block-Matching 3D-Transform (BM3D) denoiser as a new regularization term in the IR model. It is a version of the manuscript titled "Iterative Reconstruction using Block-Matching 3D-Transform (BM3D) Regularization" published in Medical Physics[30]. Section 2.2 introduces a flexible IR framework using the plug-and-play algorithm that incorporates the BM3D denoiser directly in the iterative reconstruction. It is a version of the proceeding paper titled "Iterative reconstruction using Plug-and-Play alternating direction method of multipliers (ADMM) framework" published in SPIE proceedings[31]. In Section 2.3, a Deep convolutional neural network (DCNN) was integrated into the flexible IR framework. It is a version of the proceeding paper titled "Iterative reconstruction using plug-and-play projected gradient descent" published in SPIE proceedings[32].

Chapter 3 introduces an MMD decomposition method for DECT images. It is a version of the manuscript titled "Image-domain multi-material decomposition for dual-energy CT with non-convex sparsity regularization" published in Journal of Medical Imaging[33,34].

Chapter 4 presents a series of studies on advanced treatment planning and novel delivery techniques which add new degrees of freedom to VMAT. Section 4.1 introduces the mathematical framework for VMAT, which improved the dosimetry and delivery efficiency compared with clinical VMAT algorithms. It is a version of the manuscript titled "A comprehensive formulation for Volumetric Modulated Arc Therapy," published in Medical Physics[35]. In Section 4.2, we add the dual-layer MLC to VMAT, which improves the effective modulation resolution. It is a version of the manuscript “Single-Arc VMAT optimization for Dual-Layer MLC” published on physics in medicine and biology[36]. Another way to increase the modulation resolution is to shorten the source-to-isocenter distance (SID) at the cost of limited field-of-view (FOV). In section 4.3 we developed a multi-isocenter optimization approach to compensate for the limited FOV of a single isocenter and used the group sparsity term for beam selection. It is a version of the manuscript “Many-isocenter Optimization for Robotic Radiotherapy,” published on physics in medicine and biology[37]. In the following two sections, we added more degrees of freedom to VMAT and incorporated the group sparsity term in the VMAT optimization framework for beam selection: Section 4.4 allows dynamic collimator, which increases the effective modulation resolution, and Section 4.5 allows non-coplanar trajectory that significantly expanded the VMAT searching spaces. Section 4.4 is a version of the manuscript titled “VMAT optimization with dynamic collimator rotation” published in Medical Physics[38]. Section 4.5 is a version of the manuscript titled “A novel optimization framework for VMAT with dynamic gantry couch rotation,” published in physics in medicine and biology[39]. In Section 4.6, we propose a hardware design for FLASH radiotherapy and adapt the optimization framework for the novel hardware. It is a version of the manuscript titled “ROAD: ROtational direct Aperture optimization with a Decoupled ring-collimator for FLASH radiotherapy” published in physics in medicine and biology[40].

2 USING NOVEL RECONSTRUCTION ALGORITHMS TO IMPROVE CT IMAGE QUALITY

2.1 Iterative Reconstruction using Block-Matching 3D-Transform (BM3D) Regularization

2.1.1 Introduction

Megavoltage CT (MVCT) is used in image-guided TomoTherapy treatment, but its quality is plagued by high noise levels as a result of low detector quantum efficiency (DQE) of high energy X-rays[41,42]. Both post-processing and iterative reconstruction were used to suppress the noise. In iterative CT reconstruction, regularization terms play an important role[43]. Inspired by the compressed sensing theory, L1-type regularization terms such as total variation (TV) have been widely used in CT iterative reconstructions to preserve image edges[44–46]. Recently, a tensor framelet regularization scheme, the generalization of TV, wavelet, and L1-norm, was proposed by Gao *et al.* to maintain the object boundary in MVCT reconstruction while suppressing noise[47]. These techniques are effective in removing the image noise and streaking artifacts due to view aliasing in reconstructed images but still result in noticeable image resolution loss when noise variation is high, as shown in MVCT.

The Block-Matching 3D-transform shrinkage (BM3D) algorithm was recently proposed and achieved superior image noise suppression relative to local denoising methods by clustering similar but non-local 2D image patches into one group and performing denoising within each group[48,49]. Because of the desired denoise performance, BM3D has been used in CT reconstruction. In a naïve fashion, BM3D was applied directly on the CT projection[50] or reconstructed CT images[51–53] as a preprocessing or postprocessing component separate from the reconstruction. These studies showed superior image resolution preservation to local denoising methods, but the inherent balance between data fidelity and BM3D regularization was not fully exploited as a single optimization problem. The integrated optimization approach was initially implemented for deblurring and denoising natural images[54–56], where BM3D was formulated as L1 regularization to encourage data sparsity in the BM3D transformation domain. This method exploited data consistency while suppressing image noise, better-preserved image features, and outperformed the naïve application. In 2013, Yang *et al.*[57] used BM3D patches extracted from a priori fully sampled images to regularize sparse view reconstruction of a 2D digital phantom. However, iterative CT reconstruction using BM3D on real phantom and patient projections has not been demonstrated. In this study, BM3D regularization was applied to Catphan and patient MVCT reconstruction.

2.1.2 Methods

2.1.2.1 Problem Formulation

The CT iterative reconstruction with BM3D regularization is formulated as

$$\hat{x} = \underset{x \geq 0}{\operatorname{argmin}} \frac{1}{2}\|Ax - g\|_2^2 + \beta H_{\mu}(\Phi x)\ ,$$

Equation 2-1

where g is the measured projection data, A is the forward projection matrix, Φ is the BM3D transformation matrix, β is the hyperparameter controlling the tradeoff between data fidelity and regularization, and H_{μ} is the Huber penalty function with a smoothing parameter μ[26] that approximates the L1 norm.

In Equation 2-1, the quadratic data fidelity term minimizes the discrepancy between the measured and the estimated projections, and the regularization term promotes sparsity in the BM3D transformation domain, which subsequently encourages image smoothness and maintains image texture.

As shown in Figure 2-1, the workflow involved two main steps. In the first step, a block matching process was performed on a coarsely reconstructed initial image to generate the matched groups, from which the BM3D analysis matrix Φ was constructed. In the second step, matrix Φ was utilized in Equation 2-1 for iterative reconstruction.

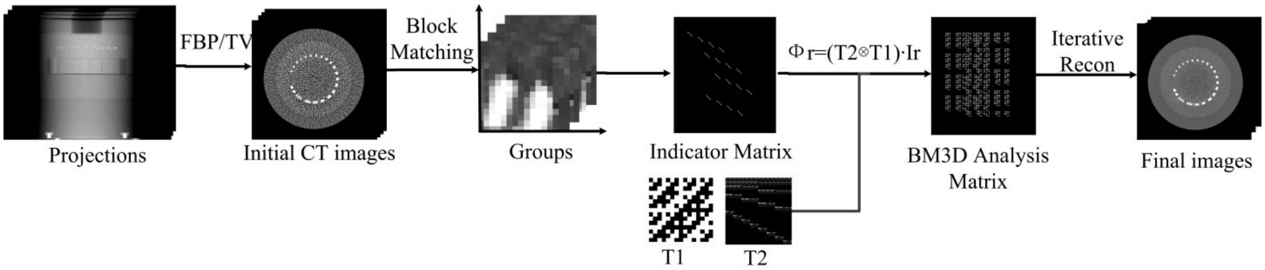


Figure 2-1 Workflow of the MVCT iterative reconstruction using BM3D regularization.

2.1.2.2 Construction of BM3D transformation matrix

The block matching was performed on the initial CT images from TV reconstruction, on which each reference patch $\mathrm{P_r}$ ($\mathrm{r} = 1,2,\cdots,\mathrm{R}$), with size $\mathrm{n^p}$ by $\mathrm{n^p}$ and a separation distance $\mathrm{n^d}$ in both the row and column directions, is denoted by the index of the voxel at the left upper corner. For each $\mathrm{P_r}$, a nonlocal searching procedure is performed within the $\mathrm{P_r}$-centered $\mathrm{n^w}$-by-$\mathrm{n^w}$ window on the initial CT image to obtain a group of similar patches $\mathrm{G_r} = \{\mathrm{S_{r,j=1}}, \mathrm{S_{r,j=2}}, \cdots, \mathrm{S_{r,j=J}}\}$ to $\mathrm{P_r}$ with respect to Euclidean distance. Let x be the vectorized MVCT image with dimension N. An indicator matrix $\mathrm{I_{r,j}}$ with dimension $(\mathrm{n^p})^2$-by-N is defined for the $\mathrm{j^{th}}$ patch in the $\mathrm{r^{th}}$ group, such that $\mathrm{I_{r,j}} \cdot \mathrm{x}$ is a vector containing every pixel in the similar patch $\mathrm{S_{r,j}}$. The indicator matrix $\mathrm{I_r}$ for group $\mathrm{G_r}$ is a concatenation of all $\mathrm{I_{r,j}}$ ($\mathrm{j} = 1,2,\cdots,\mathrm{J}$), and the matrix-vector product $\mathrm{I_r} \cdot \mathrm{x}$ provides the vectorized group $\mathrm{G_r}$.

Apart from the indicator matrix, two independent linear transformations matrices $\mathrm{T_1}$ and $\mathrm{T_2}$ are utilized in the construction of BM3D analysis matrix. $\mathrm{T_1}$ is the 1D Hadamard transformation matrix that performs interpatch transformation across different patches within one group. $\mathrm{T_2}$ is the 2D Haar wavelet matrix for intrapatch transformation within each patch. $\mathrm{T_1}$ and $\mathrm{T_2}$ together implement a 3D transformation on the group $\mathrm{G_r}$ that transforms image patches in $\mathrm{G_r}$ to the BM3D spectra, with spectra coefficients given by

$$\omega_\mathrm{r} = (\mathrm{T_1} \otimes \mathrm{T_2}) \cdot (\mathrm{I_r} \cdot \mathrm{x})$$

The joint BM3D spectra are a concatenation of the BM3D spectra for all groups. By defining the BM3D analysis matrix Φ as

$$\Phi = \begin{bmatrix} (\mathrm{T}_1 \otimes \mathrm{T}_2) \cdot \mathrm{I}_{\mathrm{r}=1} \\ (\mathrm{T}_1 \otimes \mathrm{T}_2) \cdot \mathrm{I}_{\mathrm{r}=2} \\ \vdots \\ (\mathrm{T}_1 \otimes \mathrm{T}_2) \cdot \mathrm{I}_{\mathrm{r}=\mathrm{R}} \end{bmatrix},$$

the joint BM3D spectra are related to the vectorized image x by $\omega = \Phi x$.

The objective function in Equation 2-1 was solved using FISTA. Details on the algorithm can be found in the appendix.

2.1.2.3 Evaluation

A Siemens imaging quality phantom[58], a head and neck (H&N) patient, and a prostate patient was scanned on an onboard TomoTherapy imaging system[59] with a 3.5 MV helical fan beam. A total number of 800 projections were acquired. The proposed method is compared with Filtered Back Projection (FBP), TV and BM3D post processing methods, where the BM3D post process performs the BM3D denoising filter on an FBP-reconstructed CT image. Evaluation includes both full sampled reconstruction and down sampled reconstruction, where 25% of the uniformly sampled projections are utilized. The task-based modulation transfer function (MTF)[60] is evaluated for comparison of image resolution across both linear and non-linear reconstruction methods.

2.1.3 Results

2.1.3.1 Phantom Study

The reconstruction results using BM3D regularization, TV, BM3D post process and FBP on a line pair slice and a contrast rod slice of the Siemens imaging quality phantom are shown in Figure 2-2 and Figure 2-3 respectively, for both fully sampled reconstruction and down-sampled reconstruction. The image noise, computed as the standard deviation (STD) within the squares on Figure 2-2(d1) and Figure 2-3(d1), is kept at the same level across all reconstruction methods except for FBP. TV reconstruction loses the fine feature as shown by the 8th line pair on the 800-projection image and the 7th line pair on the 200 projection image, whereas these line pairs are distinguishable on images reconstructed from other methods. BM3D post process method achieves comparably high spatial resolution as the BM3D regularization methods, as evaluated by these high contrast line pairs, but the post-processing method amplifies streaking artifacts, (Figure 2-2(c2)), which originates from the noisy FBP image in Figure 2-2(d2), due to the lack of iterative fidelity penalty in this approach. Compared to other methods, the BM3D regularization method can maintain image spatial resolution while removing both the noise and artifacts.

The contrast rod slice compares performance on high and low contrast objects across different reconstruction methods. Both TV and BM3D post process are less effective at preserving low contrast objects while removing the image noise, whereas the BM3D regularization image can distinguish low contrast fine structures that were obscured by the noise in the FBP images. Table 2-1 presents the quantitative CNR values evaluated on the contrast rod object indicated by the Region-of-Interest (ROI) with label 1 on Figure 2-3(d1).

Image resolution comparison across BM3D regularization, TV, and BM3D post process methods under the same noise level is presented by the line profile plots of the 6th line pair in Figure 2-4 and the MTF plots in Figure 2-5, evaluated basing on the highest contrast object (ROI2) on Figure 2-3(d1)). Figure 2-4 and Figure 2-5 show that, for high contrast objects, with equal image noise suppression, both the BM3D post process and the BM3D regularization method maintained resolution comparable to the FBP, whereas the TV regularization degraded the resolution. At 50% of the MTF magnitude, the spatial resolution using BM3D regularization is 5.00 lp/cm and 4.33 lp/cm, whereas that of using TV is 1.79 lp/cm and 1.70 lp/cm, for the fully sampled reconstruction and down-sampled reconstruction, respectively.

Figure 2-6 shows the tradeoff between image noise and resolution for different reconstruction methods. The image noise is computed as the STD of the square on Figure 2-3(d1), and the image resolution is calculated as the area under the MTF curve, where the task-based MTF is evaluated on the object in ROI2 on Figure 2-3(d1). Across a suitable range of image noise, both the BM3D regularization and the BM3D post process were able to maintain a comparably high image resolution compared with TV, for both the fully sampled reconstruction and down-sampled reconstruction. The images shown in Figure 2-3 corresponds to the noise STD at 0.14 and 0.18 for 800 projections and 200 projections, respectively.

CNR	BM3D regularization	TV	BM3D Post Processing	FBP
800 projections	372.10	222.10	300.61	48.91
200 projections	235.09	230.75	203.34	14.47

Table 2-1 The CNR for the contrast slice with the ROI1 indicated in Figure 2-3(d1).

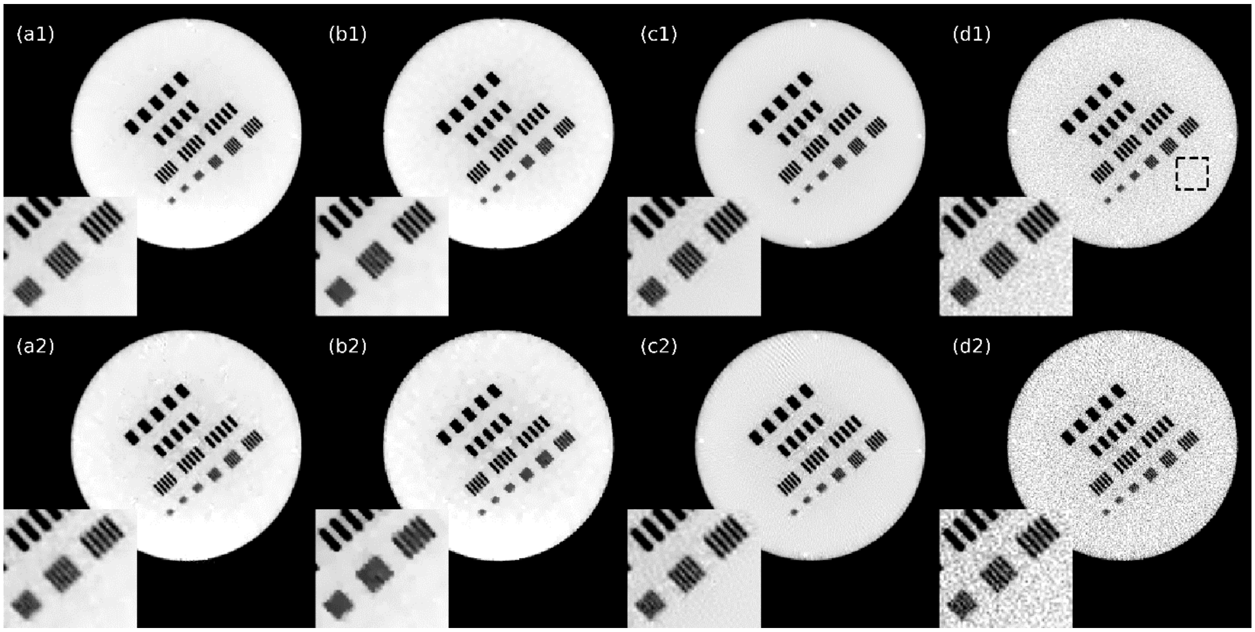


Figure 2-2 Resolution slice reconstructed from (a) BM3D regularization, (b) TV, (c) BM3D post process, and (d) FBP using (1) 800 projections and (2) 200 projections, respectively. Zoom-in details for the 6th—8th line pairs are shown in the left lower corner. The images from BM3D regularization, TV, and BM3D Post Process are under the same noise level.

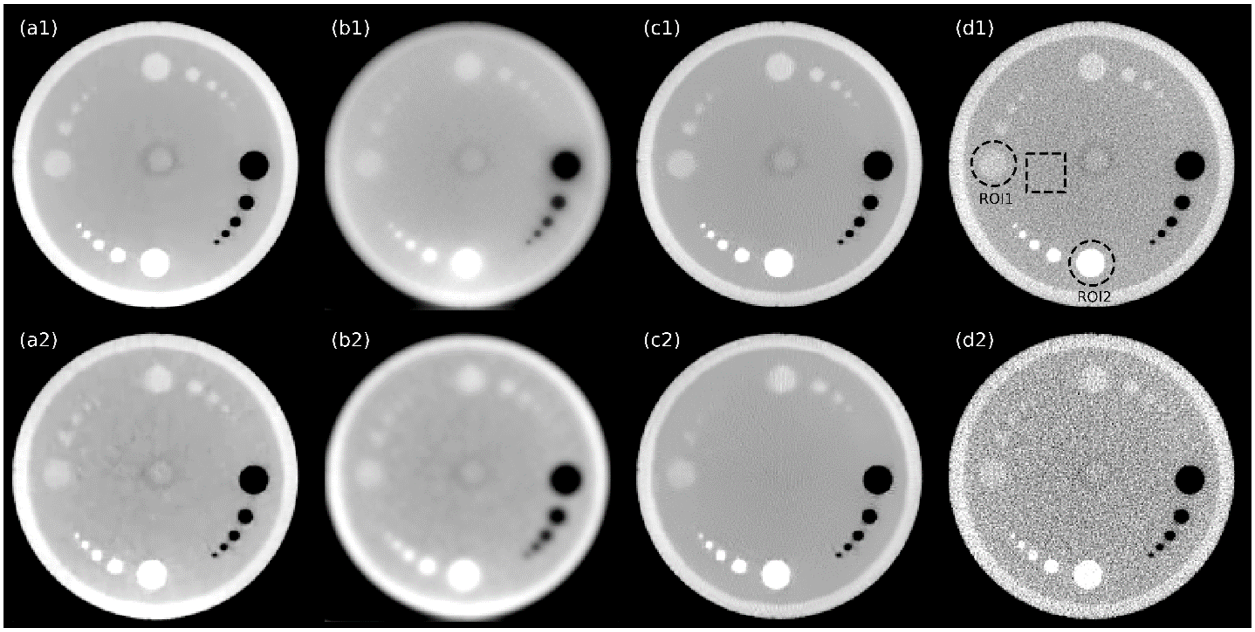


Figure 2-3 Contrast rod slice reconstructed from (a) BM3D regularization, (b) TV, (c) BM3D post process, and (d) FBP using (1) 800 projections and (2) 200 projections, respectively. The images from BM3D regularization, TV, and BM3D Post Process are under the same noise level.

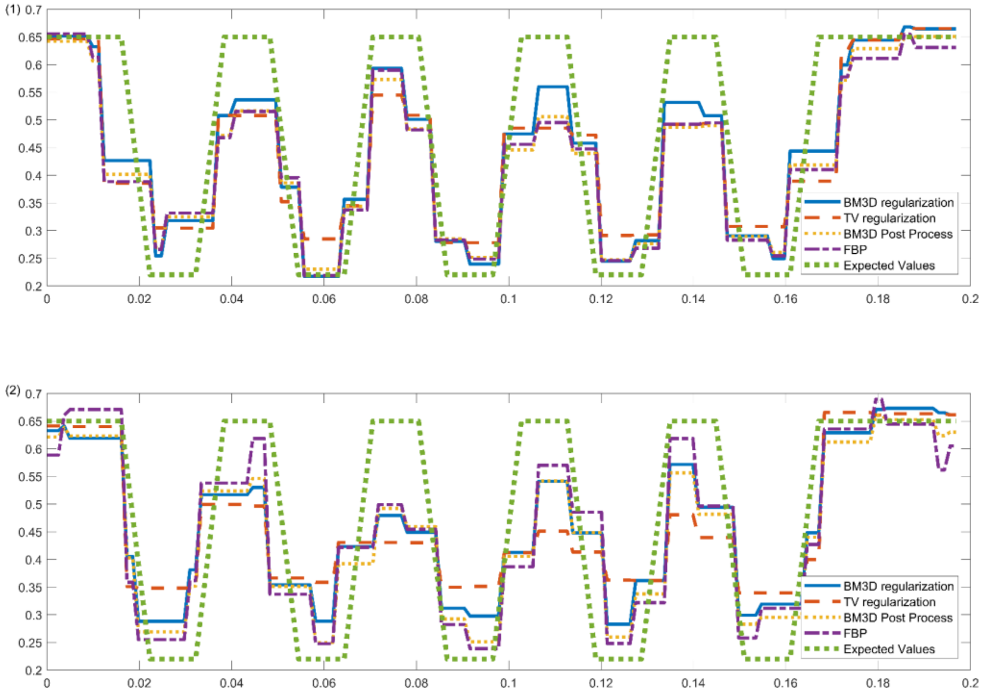


Figure 2-4 Crossline plots of the 6th line pair on the resolution slice for all methods and the expected values, using (1) 800 projections and (2) 200 projections. The images from BM3D regularization, TV, and BM3D Post Process are under the same noise level. The unit of x axis is cm, and the unit of y axis is cm^{-1}.

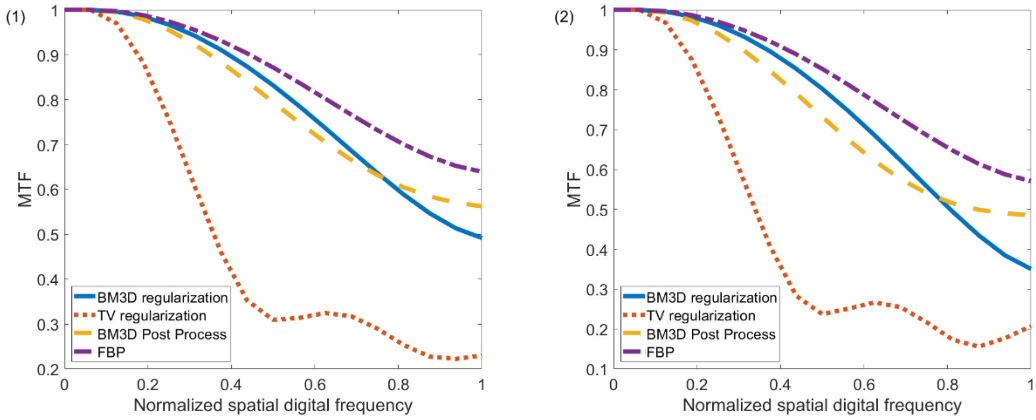


Figure 2-5 MTF of the MVCT images reconstructed from BM3D regularization, TV, BM3D Post Process, and FBP using (1) 800 projections and (2) 200 projections. The images from BM3D regularization, TV, and BM3D Post Process are under the same noise level.

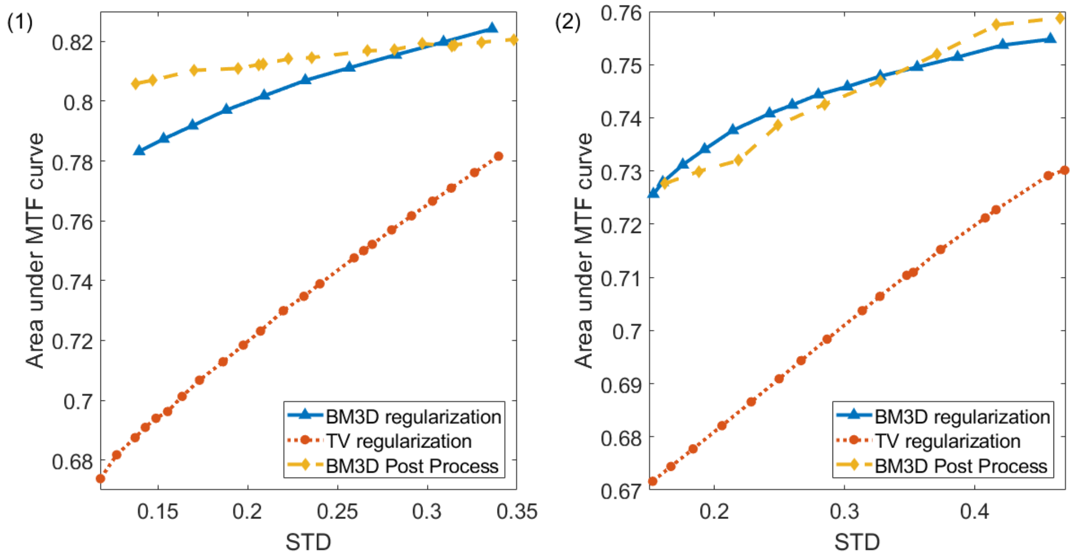


Figure 2-6 The resolution vs noise curve of BM3D regularization, TV, and BM3D Post Process using (1) 800 projections and (2) 200 projections. The resolution is computed as the area under the MTF curve, evaluated on the ROI2 on Figure 2-3 (d1), and the image noise is computed as the STD of the square on Figure 2-3 (d1).

2.1.3.2 Patient Study

The reconstruction results for a H&N patient slice and a prostate patient slice using BM3D regularization, TV, BM3D post-process, and FBP are shown in Figure 2-7 and Figure 2-8, respectively, for both fully sampled reconstruction and down-sampled reconstruction. The image noise, computed as the STDs of the squares on Figure 2-7 (d1) and Figure 2-8(d1), is kept at the same level across all reconstruction methods except for FBP. TV images lose fine bone structures, as shown in the zoom-in ROI displays of a H&N patient in Figure 2-7, and produce cartoon-like artifacts characteristic of TV regularization. BM3D post-process images can preserve image spatial resolutions in the high-contrast region comparable to BM3D regularization and FBP but are less effective in preserving the low-contrast structures, shown in the zoom-in ROIs. For the prostate patient, the TV images showed more blurred bony boundaries. Compared with BM3D regularization, the BM3D post process results in worse soft tissue contrast, as shown in Figure 2-8(c2), and magnifies the FBP image artifacts.

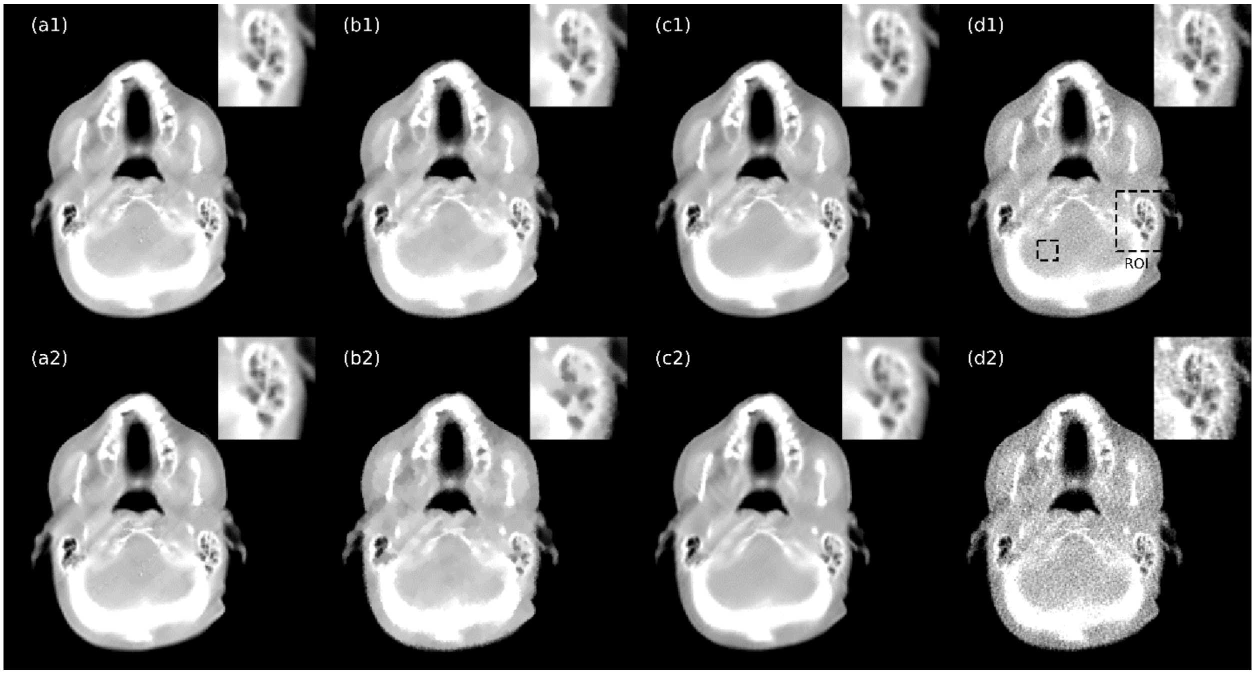


Figure 2-7 H&N images with zoom-in displays of ROI reconstructed from (a) BM3D regularization, (b) TV, (c) BM3D post process, and (d) FBP using (1) 800 projections and (2) 200 projections, respectively.

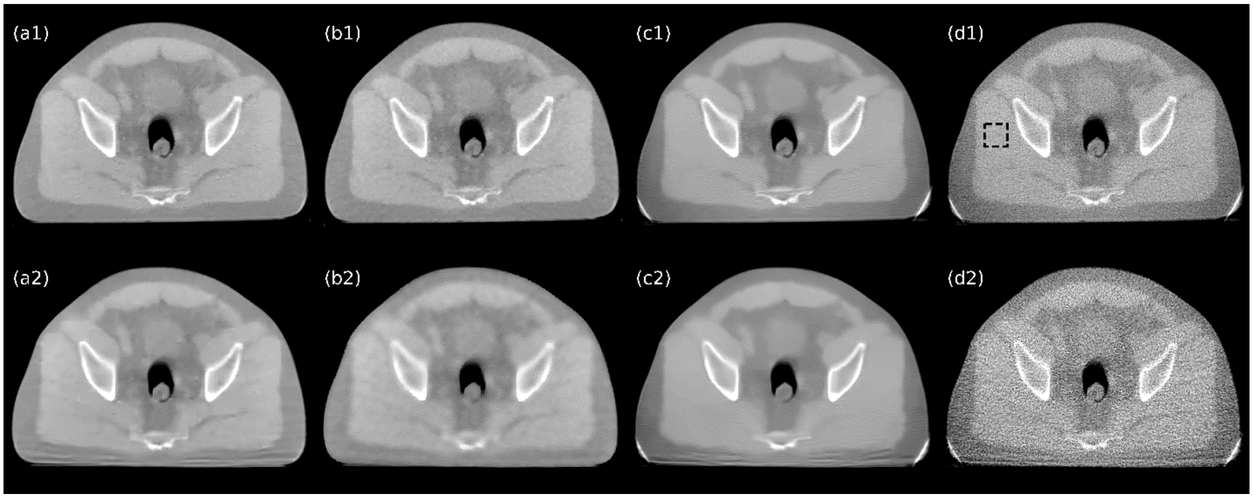


Figure 2-8 Prostate images reconstructed from (a) BM3D regularization, (b) TV, (c) BM3D post process, and (d) FBP using (1) 800 projections and (2) 200 projections, respectively.

2.1.4 Discussion

CT imaging dose is a concern for diagnostic imaging and image-guided radiation therapy, motivating the development of low dose CT. In practice, imaging dose reduction is achieved using sparse projection sampling and/or reduced x-ray tube current[10,11]. Due to the ill-conditioned system matrix brought by the heavily undersampled data and/or high noise in low-dose CT imaging, iterative reconstruction has shown superiority in image artifact reduction and noise control in comparison to the analytical filtered-backprojection (FBP) algorithm since the former incorporates physical constraints and image features into the iterative framework as regularization terms[43]. Because of the low DQE, MVCT can be a viewed as a special case of low dose CT that can benefit from iterative CT reconstruction.

In this work, we for the first time implemented non-local BM3D regularization as L1 regularization for raw CT data reconstruction. Compared with TV regularization, both BM3D post process and regularization are more effective in maintaining the resolution while reducing the noise. BM3D regularization, however, is better at enhancing low contrast conspicuity and controlling artifacts than the BM3D post-process due to the iterative application of the fidelity term.

One limitation of the BM3D regularization is its higher computational cost. Compared with the 2D natural image processing reported by Danielyan et al.[54], the CT reconstruction problem is several orders of magnitude greater. The previous method of solving BM3D regularized iterative reconstruction using ADMM[57], except for the difference of relying on a priori data for reconstruction, is impractically slow for real patient data. In this study, to improve the computational speed, the optimization problem was solved using FISTA[27], a fast proximal gradient method that significantly reduces the computation cost. In addition, to reduce matrix multiplication time and reduce memory usage, the BM3D analysis matrix was implemented as sequential multiplication of small transformation matrices in different dimensions[48]. With these acceleration methods, reconstruction for one image slice currently takes approximately 30 seconds with BM3D regularization, whereas FBP reconstruction and TV reconstruction takes less than 0.2s and 3 seconds, respectively, on an Intel Core i7-7700K CPU with 64 GB RAM and a GTX TITAN X. As a future development, the embarrassingly parallelizable matrix multiplication time can be further reduced with hardware acceleration techniques, e.g., graphics processing unit (GPU)[61].

2.1.5 Summary

As the first step to incorporate novel image denoisers into CT iterative reconstruction (CTIR), in this study, we modified and reformulated the BM3D denoiser as the BM3D regularization term in the IR model. The integrated model is simple and concise in the form of a convex optimization problem. Yet, it includes the non-local patch-wise feature originated from BM3D denoisers, which contributed to its improved image SNR and image resolution compared with conventional IR models with TV regularization. On the other hand, the proposed reformulation is unique to BM3D denoiser, and the method is not compatible with or adaptable to many other state-of-the-art image denoisers. The following sections overcome this limitation by introducing a more flexible integration method that can be applied to almost any denoisers.

2.2 Iterative reconstruction using Plug-and-Play alternating direction method of multipliers (ADMM) framework

2.2.1 Introduction

Low dose CT (LdCT) has been investigated to reduce the associated radiation risks of diagnostic CT. In practice, low-dose data acquisition is typically achieved using sparse projection sampling and/or reducing x-ray tube current. Due to heavily under-sampled data acquisition and/or low detector photon counting, the system matrix is highly ill-posed, resulting in severe image artifacts and overwhelming noise using the conventional analytical FBP algorithm. Sparse regularized iterative reconstruction models have been investigated to mitigate image artifacts and noise. The iterative reconstruction significantly enhanced the image signal-to-noise ratio compared with analytical FBP but often compromised the image resolution and low-contrast object visibility [62].

Recent emerging denoising techniques such as the BM3D[49] technique has been shown to remove image noise while better retaining image details. Applying the BM3D denoiser on FBP reconstructed CT image, termed BM3D post-process method, has shown to improve soft tissue conspicuity[51]. However, incorporating these state-of-the-art denoisers into model-based reconstruction remains challenging since these denoising methods do not have a known formulation as an optimization problem.

The previous section introduces a BM3D regularized iterative reconstruction method for MVCT[63], which utilizes a non-local patch-wise regularization term to promote sparsity in the BM3D transformation domain. This method effectively reduced image noise and showed improved low-contrast conspicuity compared with the BM3D post-process method and conventional model-based method such as TV. However, there are three intrinsic limitations with the BM3D regularization method. Firstly, the BM3D regularization reconstructed image is highly correlated with an initial input image, either from FBP or TV, which was utilized in the block-matching procedure to build up the BM3D transformation matrix. Since the initial images may be noisy or over-smoothed, a single-step block-matching could impose a bias and amplify patterned artifacts on the reconstructed image. The problem is avoided in the original BM3D denoiser in [49] through two denoising steps. The first step performs a basic hard-thresholding approach to estimate a denoised image, based on which the block-matching was recalculated for the second step. Secondly, the BM3D regularization applies a uniform soft-thresholding on every element, whereas the original BM3D denoiser utilizes an empirical Wiener filter that performs element-by-element filtering based on the denoised image from the first step. It has been observed in the experiments that the Wiener filtering restores more details and improves the denoising performance [48]. Thirdly, the patch-wise BM3D regularization tends to induce patch-wise image artifacts, especially in highly under-sampled data acquisition. To reduce patch-wise artifacts and border effects, the original BM3D denoiser utilized the Kaiser window in combination with an adaptive aggregation that gives priority to homogeneous patches. These features were not incorporated in the BM3D regularization method due to the difficulty of formulating them into mathematical terms.

A flexible Plug-and-Play (PnP) framework was proposed in [64] that allows to plug in any off-the-shelf denoisers to replace a module in the alternating direction method of multipliers (ADMM) algorithm [28] for model-based image denoising problem. The PnP ADMM algorithm does not require formulating the denoiser into an optimization problem. Instead, the denoiser is applied directly as an iteration step. Studies [64–67] have shown promising empirical results on Gaussian and Poissonian image restoration problems using the PnP ADMM framework. In this study, the PnP ADMM framework was developed for model-based CT reconstruction to fully exploit the data consistency while using the state-of-art BM3D denoiser to suppress the image noise.

2.2.2 Method

2.2.2.1 ADMM for CT iterative reconstruction

A typical CT iterative reconstruction framework is formulated as

$$\hat{x} = \underset{x \geq 0}{\operatorname{argmin}} \frac{1}{2} \|Ax - b\|_2^2 + \beta R(x),$$

Equation 2-2

where b is the measured projection data, A is the forward projection matrix. The first term is the quadratic data fidelity term that minimizes the discrepancy between the measured and the estimated projections. $R(x)$ is the regularization term that incorporates prior information on the image, β is the hyperparameter controlling the tradeoff between data fidelity and regularization.

ADMM is one of the first-order algorithms that is well suited to large-scale convex optimization problems. The algorithm solves the optimization problem of the form [28]

$$\underset{x,y}{\text{minimize}} \quad f(x) + g(y)$$

$$\text{subject to} \quad Cx + Dy = c,$$

Equation 2-3

where f and g are convex functions, and A, B are linear operators. Note that the objective function is split into two parts, each part has one optimization variable. The augmented Lagrangian of the optimization problem is defined as

$$L_\rho(x, y, v) = f(x) + g(y) + v^T(Cx + Dy - c) + \frac{\rho}{2}\|Cx + Dy - c\|_2^2.$$

Equation 2-4

ADMM solves the optimization problem in Equation 2-3 by minimizing the augmented Lagrangian in Equation 2-4 with respect to the optimization variables x and y and updating the dual variable v in each iteration through the following equations:

$$x^{k+1} := \underset{x}{\operatorname{argmin}}\ L_\rho(x, y^k, v^k)$$

$$y^{k+1} := \underset{y}{\operatorname{argmin}}\ L_\rho(x^{k+1}, y, v^k)$$

$$v^{k+1} := v^k + \rho(Cx^{k+1} + Dy^{k+1} - c).$$

Equation 2-5

The CT iterative reconstruction problem can be formulated into the canonical form of ADMM shown in Equation 2-3 by defining

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2$$

$$g(y) = \beta R(y) + I_+(y)$$

$$C = I, D = -I, c = 0,$$

where $I_+(y)$ is an indicator function imposing the non-negativity constraint

$$I_+(y) = \begin{cases} 0 & if\ y \geq 0 \\ \infty & otherwise. \end{cases}$$

Subsequently, the ADMM iterations in Equation 2-5 are adapted for the CT iterative reconstruction problem as:

$$x^{k+1} \coloneqq \underset{x}{\operatorname{argmin}} \ \frac{1}{2}\|Ax-b\|_2^2 + \frac{\rho}{2}\|x-\tilde{x}^k\|^2$$

$$y^{k+1} \coloneqq \underset{y \geq 0}{\operatorname{argmin}} \ \beta R(y) + \frac{\rho}{2}\|y-\tilde{y}^k\|^2$$

$$v^{k+1} \coloneqq v^k + \rho(x^{k+1} - y^{k+1})$$

Equation 2-6

where $\tilde{x}^k = y^k - v^k/\rho$, and $\tilde{y}^k = x^{k+1} + v^k/\rho$.

2.2.2.2 Plug-and-Play with ADMM

Conjugate Gradient algorithm: Solving $\boldsymbol{Wx = d}$

```
Initialization: $x \coloneqq 0,\ r \coloneqq d,\ \rho_0 \coloneqq \|r\|_2^2$
For $k = 1, 2, \ldots$ do
    if $k = 1$
        $p \coloneqq r$
    else
        $p \coloneqq r + (\rho_{k-1}/\rho_{k-2})p$
    end
    $u \coloneqq Wp$
    $a \coloneqq \rho_{k-1}/p^T u$
    $x \coloneqq x + ap$
    $r \coloneqq r - au$
    $\rho_k \coloneqq \|r\|_2^2$
end
```

Table 2-2 Conjugate Gradient algorithm: Solving $\boldsymbol{Wx = d}$

The first subproblem with respect to variable x in Equation 2-6 is a least square problem that is equivalent to a linear equation

$$(A^T A + \rho I)x^{k+1} = A^T b + \rho\tilde{x}^k.$$

Equation 2-7

Due to the size of the system, it is infeasible to solve Equation 2-7 via matrix inversion. Instead, it was solved iteratively through the Conjugate Gradient method [68], which is one of the most popular and efficient iterative methods for solving sparse symmetric positive definite (SPD) systems of linear equations. The pseudo code of the CG algorithm following C.T.Kelley [69] is presented in Table 2-2.

```
Plug-and-Play ADMM
Initialization: $\rho_0 := 100,\ \eta < 1,\ \gamma > 1$
For $k = 1, 2, \ldots$ do
    $x_{k+1} := \underset{x}{\text{argmin}}\ \frac{1}{2}\|Ax - b\|_2^2 + \frac{\rho_k}{2}\|x - (y_k - z_k)\|_2^2$
    $\sigma_k := \sqrt{\frac{\beta}{\rho_k}}$
    $y_{k+1} := D_{\sigma_k}(x_{k+1} + z_k)$
    $z_{k+1} := z_k + x_{k+1} - y_{k+1}$
    $\Delta_{k+1} := \frac{1}{\sqrt{n}}(\|x_{k+1} - x_k\|_2 + \|y_{k+1} - y_k\|_2 + \|z_{k+1} - z_k\|_2)$
    if $\Delta_{k+1} \geq \eta\Delta_k$
        $\rho_{k+1} := \gamma\rho_k$
    else
        $\rho_{k+1} := \rho_k$
    end
end
```

Table 2-3 Plug-and-Play ADMM

The second sub-problem with respect to variable y includes a least square data fidelity term and a regularization term $R(y)$:

$$y^{k+1} := \underset{y}{\text{argmin}}\ R(y) + \frac{1}{2\sigma^2}\|y - \tilde{y}^k\|^2,$$

Equation 2-8

where $\sigma = \sqrt{\beta/\rho}$. Note that it is equivalent to a denoising problem with $R(y)$ as prior, $\tilde{y}^k$ as the noisy input, and σ as the estimated noise level. The optimum y^{k+1} could be viewed as a cleaned version of the input $\tilde{y}^k$. In the PnP framework, instead of solving the optimization problem in Equation 2-8, it utilizes an off-the-shelf image denoiser, denoted by D_σ, to derive the cleaned version of the input as the optimum y^{k+1}. At each iteration, the image denoiser is plugged in and the cleaned image y^{k+1} is subsequently utilized to update x and z. In this study, BM3D is chosen as the off-the-shelf image denoiser, where the hyperparameter σ

indicates the estimated noise level of the acquisition data and controls the amount of smoothing in the hard-thresholding step and the Wiener filtering step. The pseudo code of the PnP algorithm is summarized in Table 2-3.

2.2.2.3 Evaluation

The proposed PnP ADMM was evaluated on low dose scans of CT ACR 464 phantom and two lung screening scans and is compared with FBP, Total Variation (TV), BM3D post-process, and the BM3D regularization. The reconstruction utilized the FreeCT wFBP GPU code in [70] for FBP and a GPU-based separable footprint forward and backward projector in [71] for iterative reconstruction. FBP reconstructed images using standard dose scan are provided as the ground truth. All reconstruction methods besides 'FBP standard dose' were evaluated on low dose data having 5.5%, 50%, and 25% of the standard dose in the ACR phantom, lung #1, and lung #2 cases, respectively.

For the ACR phantom, we acquired raw projection data with standard-dose protocol and low dose protocol on Siemens Definition without flying focal spot. For the two lung screening cases, raw projection data with standard-dose protocol was acquired on Siemens Definition with the z-flying focal spot, and dose reduction was simulated using a previously validated tool [72,73] that takes into account Poisson statistics, the scanner bowtie, and measurements of the real scanner electronic noise.

2.2.3 Results

Figure 2-9 is the reconstructed image with zoomed-in ROI of the ACR phantom using PnP ADMM with BM3D, TV, BM3D regularization, BM3D post-process, and FBP, from low dose

CT scans, along with FBP from standard-dose CT scans. The ROI showing the 6th line pairs at 9 lp/cm resolution is indicated by the red square. With reduced image noise, the PnP ADMM (Figure 2-9A) clearly distinguished the 9 lp/cm line pairs and achieved a higher image resolution than the FBP (Figure 2-9 E), from low dose CT data. The 9 lp/cm line pairs are indistinguishable using other iterative reconstruction methods, including TV (Figure 2-9B) and BM3D regularization (Figure 2-9C) under the same noise level. The BM3D post process method removed the image noise at the cost of blurred line pairs (Figure 2-9D). To the best of our knowledge, the BM3D filter is a leading image denoiser particularly strong at maintaining structured image patterns, such as the line pairs in the ACR phantom image. Recovering such resolution may not be possible using other denoising filters. More remarkably, the line pairs are better recovered in the PnP ADMM image (Figure 2-9A) with 5.5% of the imaging dose than the FBP image from standard-dose scans (Figure 2-9F).

Figure 2-10 shows the line profile of the yellow line shown on Figure 2-9A that crosses the 9 lp/cm line pairs. The expected oscillatory pattern could be identified on the line profile for the FBP standard dose image and the PnP ADMM low dose image, with the latter achieving substantially higher contrast. The line pair patterns are completely blurred or lost to noise with other methods, consistent with the visual examination of the reconstructed images in Figure 2-9.

Figure 2-11 shows the reconstructed image of lung patient #1 with a zoomed-in view of the fissure line in the left lung. The PnP ADMM (Figure 2-11A) reduced the image noise and kept most image details, and the fissure line is clearly visible. The total variation (Figure 2-11B) and BM3D post-process (Figure 2-11D) lost fine details and blurred the low contrast structures to achieve the same reduced noise level, consistent with our previous study [63].

The BM3D regularization method (Figure 2-11C) achieved fair image resolution, but it also induced patch-wise artifacts. The fissure line is visible on the FBP images (Figure 2-11E and Figure 2-11F), but was contaminated by noise, especially with the low dose scan.

Figure 2-12 shows the reconstructed image of lung patient #2 with a zoomed-in view of the fissure line in the left lung. The PnP ADMM (Figure 2-12A) shows the fissure line most clearly, while other iterative reconstruction methods (Figure 2-12B, Figure 2-12C) or smooth filter (Figure 2-12D) present the fissure line with reduced contrast under the same image noise level. The BM3D regularization method (Figure 2-12C) also presents patch-wise noise, leading to conspicuity problems within the noisy patches. Both FBP low dose image (Figure 2-12E) and FBP standard dose image (Figure 2-12F) are noisier than other methods.

Table 2-4 summarizes the mean and standard deviation (STD) of a conformal region, indicated by the blue square on the reconstructed images (Figure 2-9, Figure 2-11, and Figure 2-12) of the ACR phantom and two lung patients, respectively. The noise levels are comparable across iterative reconstruction methods and the BM3D post-process method for a fair comparison of image resolution.

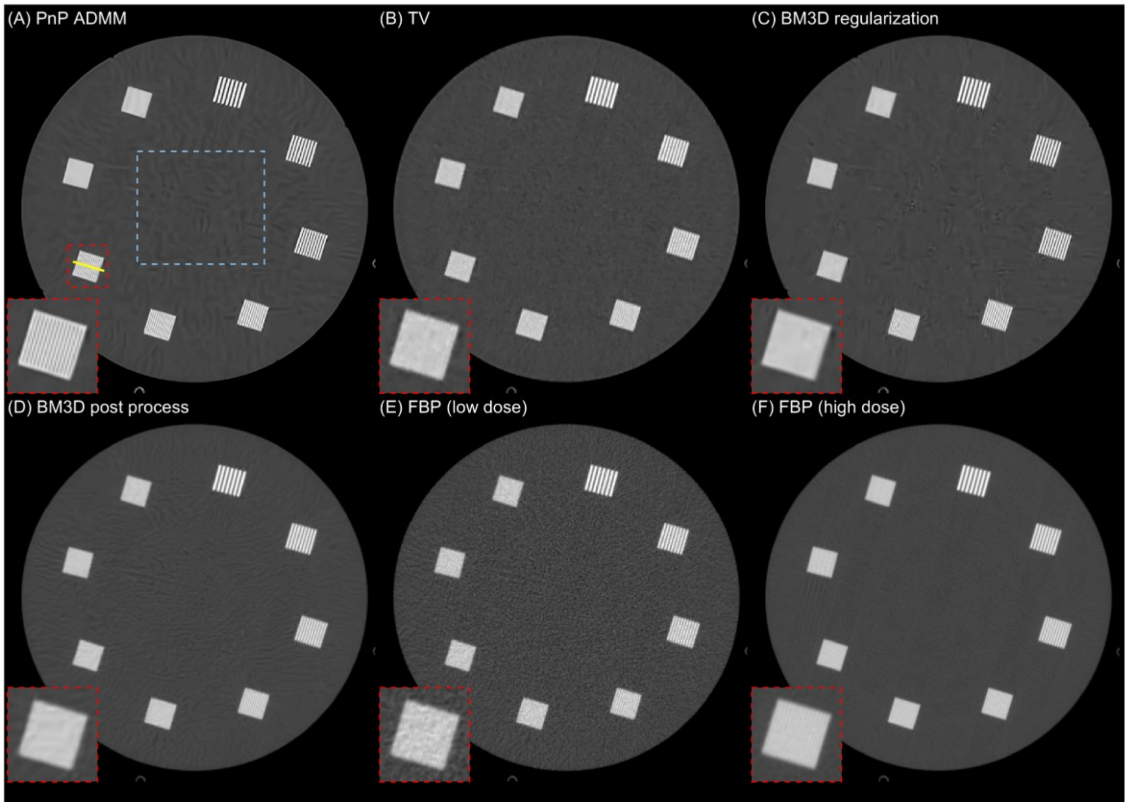


Figure 2-9 Reconstructed image with zoomed in ROI (red square) of the ACR phantom using (A) Plug-and-Play ADMM, (B) TV, (C) BM3D regularization, (D) BM3D post process and (E) FBP from low dose CT data, and (F) FBP from standard dose CT data. The conformal region indicated by the blue rectangle is used for standard deviation evaluation in Table 2-4. The line profile of the yellow line on Figure 2-9(A) is shown in Figure 2-10.

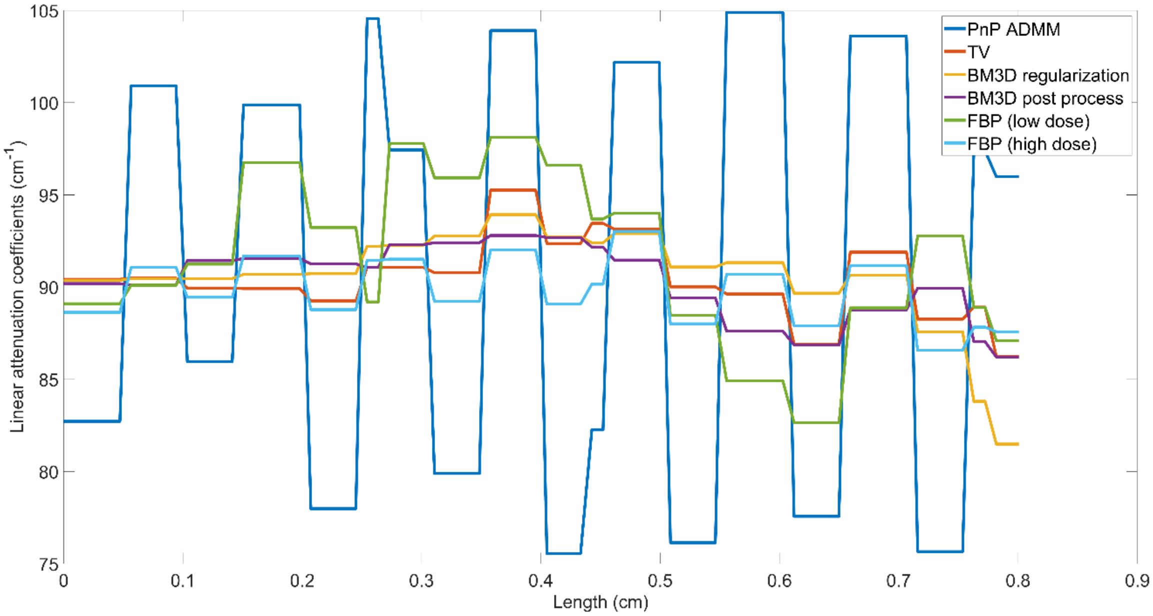


Figure 2-10 Line profile across all methods for the ROIs on Figure 2-9.

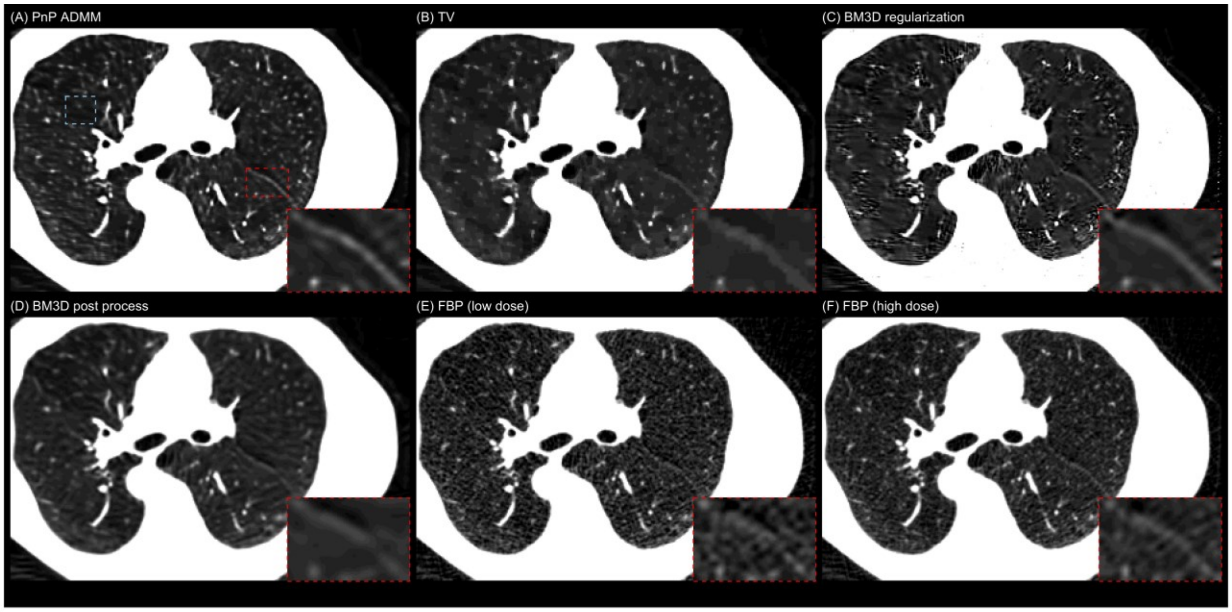


Figure 2-11 Reconstructed image of the lung patient #1 with zoomed-in view of the fissure line using (A) Plug-and-Play ADMM, (B) TV, (C) BM3D regularization, (D) BM3D post process and (E) FBP from low dose CT (50% dose), and (F) FBP from standard dose CT data. The conformal region indicated by the blue rectangle is used for standard deviation evaluation in Table 2-4.

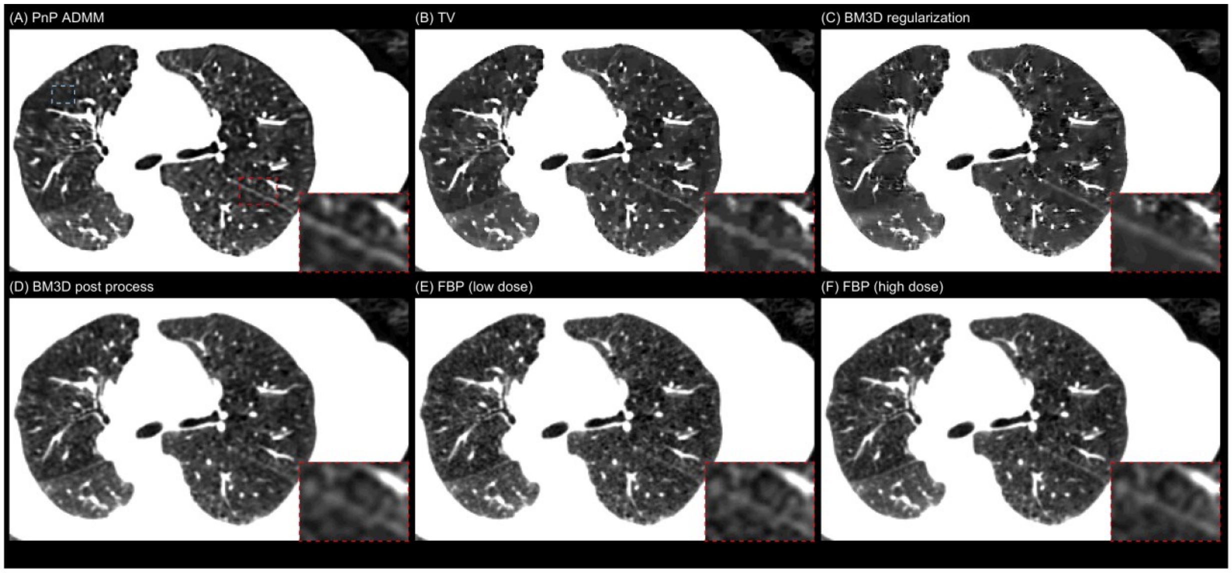


Figure 2-12 Reconstructed image of the lung patient #2 with zoomed-in view of the fissure line using (A) Plug-and-Play ADMM, (B) TV, (C) BM3D regularization, (D) BM3D post process and (E) FBP from low dose CT data (25% dose), and (F) FBP from standard dose CT data. The conformal region indicated by the blue rectangle is used for standard deviation evaluation in Table 2-4.

	PnP ADMM	**TV**	**BM3D regularization**	**BM3D post-process**	**FBP low dose**	**FBP standard dose**
ACR	79±27	79±28	78±27	79±27	79±84	78±23
Lung #1	-866±53	-856±42	-858±50	-854±51	-852±108	-852±87
Lung #2	-897±24	-902±25	-904±27	-898±25	-899±35	-897±30

Table 2-4 The mean and STD of all methods evaluated on the conformal region, indicated by the blue square on the reconstructed images (Figure 2-9, Figure 2-11, and Figure 2-12) of the ACR phantom and two lung patients, respectively.

2.2.4 Discussions

The previous study in Section 2.1 adapted BM3D denoiser into CT iterative reconstruction by constructing non-local patch-wise regularization based on the block-matching procedure on the initial input image. The non-local BM3D regularization improved image resolution compared with TV under the same noise level, but the image quality is susceptible to the initial input image-induced patch-wise artifacts due to the difficulty of rigorously formulating the empirical artifacts-removal procedures in BM3D into the optimization problem. The proposed PnP framework incorporates all features of the original BM3D filter into the model-based image reconstruction framework. Therefore it not only inherits the rigorous mathematical framework of model-based reconstruction from highly under-sampled data acquisition but also fully takes advantage of the BM3D denoiser on resolving image details while removing noises.

In this study, BM3D denoiser is used as the plug-in component, but the PnP framework is not limited to BM3D. For example, studies have shown improved image quality of incorporating other image denoisers into the PnP framework, including non-local means [65], Gaussian mixture models [65], and Deep convolutional neural network (DCNN) [74], compared with applying these denoisers directly on the image.

The current PnP framework is based on the ADMM algorithm, which requires solving a linear equation involving the system matrix at each iteration. While it is trivial to solve the linear equation for many image restoration problems where the multiplication with the system matrix and its transpose reduces to element-wise multiplication in the frequency domain under Fast Fourier Transform (FFT), it is difficult to evaluate the same equation for CT reconstruction problem. Iterative optimization algorithms such as the CG algorithm are required for solving the linear equation at each ADMM iteration. Even with the acceleration of the projection and back-projection process through graphic processing units (GPUs) [75], the PnP-ADMM still takes around 2 hours on a single desktop with 4 TITAN X GPUs, while a standard iterative reconstruction using TV takes around 15 minutes on the same computer and for the same case. Studies [76,77] have investigated applying the PnP framework to other first-ordered algorithms such as Primal-dual hybrid gradient method (PDHG)[78] and FISTA [27], which only requires multiplication with the system matrix at each iteration and therefore substantially reduces the computational cost for a large-scale system. The potential of using these algorithms to reduce the reconstruction time of PnP will be investigated in future work.

2.2.5 Summary

Here we introduce a novel flexible CTIR model using the plug-and-play algorithm, where an off-the-shelf denoiser, the BM3D denoiser, can be plugged into one step of the iterative computation. Compared with the previous BM3D regularization method, which modifies the BM3D denoiser to be compatible with IR, the PnP algorithm in this study uses the original BM3D denoiser and demonstrates improved low-contrast conspicuity, higher image resolution, and reduced image artifacts compared. In addition, the PnP algorithm is not limited to the BM3D denoiser, as will be demonstrated in Section 2.3.

2.3 Iterative reconstruction using plug-and-play projected gradient descent

2.3.1 Introduction

Cone-beam breast CT (CBBCT) has the potential to overcome the limitations of mammography, such as overlapping of breast tissue, breast compression, and low detectability in dense breast tissue. However, the additional radiation dose hinders the application of CBBCT to a large population, which motivated the development of low-dose CBBCT (LdCBBCT).

Emerging denoising techniques such as the BM3D[49] and DCNN have shown improved image resolution, and soft tissue conspicuity through post-processing on the FBP reconstructed images [51,79]. However, these post-processing methods heavily rely on the initial FBP image and lack of robustness for the extremely noisy LdCBBCT reconstruction.

To overcome the intrinsic limitations of FBP under the ultra-low-dose scenario, studies have investigated combining state-of-the-art denoisers with model-based reconstruction for low dose fan-beam CT (LdFBCT). In [80], the BM3D denoiser was simplified and formulated as a regularization term in the model-based reconstruction. In the plug-and-play (PnP) alternating direction method of multiplier (ADMM) framework [31], the BM3D denoiser replaced a module in the ADMM for model-based reconstruction [28]. In [81], the CNN-based projected gradient descent method combines the fbpconvnet [79] with the proximal gradient descent (PGD) algorithm for LdFBCT.

Compared with LdFBCT, LdCBBCT reconstruction is more challenging due to high noise, high computation cost, and lack of training data. In this study, we adapted the PnP-PGD framework proposed in [81] for model-based CBBCT reconstruction with the following considerations: 1. The framework incorporates fbpconvnet into iterative reconstruction to fully exploit the data consistency while effectively suppressing the image noise. 2. fbpconvnet can be trained with low/high dose diagnostic CT images other than the CBBCT data, and the data consistency is maintained by minimizing the data fidelity term. 3. The proximal gradient descent only requires forward and backward projection once in each iteration, substantially reducing the computational cost compared with ADMM.

2.3.2 Method

2.3.2.1 PGD for CT iterative reconstruction

PGD solves the optimization problem in Equation 2-2 via the following fixed-point iteration:

$$x^{k+1} \coloneqq \mathrm{Prox}_{tR}\left(x^k - tA^T(Ax^k - g)\right),$$

Equation 2-9

where Prox_{tR} is the proximal operator[26]. Notice that the minimization problem with respect to the optimization variable y is equivalent to a denoising problem with $R(y)$ as prior and x as the noisy input. Subsequently, the optimum $\hat{y}$ can be viewed as a denoised version of the noisy input x. In the PnP framework, an off-the-shelf image denoiser, denoted by D_σ, is used to replace the sub-problem optimization. The method is termed plug-and-play proximal gradient descent method (PnP-PGD):

$$x^{k+1} := D_{\sigma}\left(x^k - tA^T(Ax^k - g)\right).$$

Equation 2-10

In this study, the fbpconvnet [79] is used as the off-the-shelf image denoiser, which is one of the most popular deep CNN denoisers for low dose CT.

2.3.2.2 Data preparation, training, and evaluation

Due to limited breast CT datasets, the fbpconvnet was trained using low and standard dose FBCT. The standard dose FBCT projection data were acquired on Siemens Definition with the z-flying focal spot. The low dose FBCT projection data was simulated using a previously validated tool [72,73], which takes into account Poisson statistics, the scanner bowtie, and measurements of the real scanner electronic noise. A total of 3600 FBP-reconstructed [70], low and high dose FBCT images from six lung screening patients were utilized for the training. The network structure was described in a previous publication [79].

The proposed PnP-PGD with fbpconvnet (PnP-fbpconvnet) was compared with FBP, TV, BM3D post process, and fbpconvnet post-process method, using low dose, medium dose, and standard dose CBCT scans of a breast phantom. The low (or medium) dose scans were achieved by placing a 2mm (or 1mm) aluminum plates in front of the x-ray tube to attenuate the beam. Assuming the mean energy of the x-ray beam is at 40 keV, the low and medium dose scans correspond to around 70% and 85% dose compared with the standard dose scans. A GPU based forward and backward projector in [82] was used for CBCT iterative reconstruction. FBP reconstructed images using standard dose scans are provided as ground truth.

2.3.3 Results

Figure 2-13 shows the reconstructed images from medium dose scans (1mm Al). All denoising techniques significantly reduced image noise compared with FBP-0mm and FBP-1mm. The BM3D post-processing method results contain substantial streaking artifacts due to its inability to differentiate artifacts from image structures. Total variation image shows the 'staircase' phenomenon and streak noise. Although the fbpconvnet was trained with LdFBCT and applied on LdCBBCT, both the fbpconvnet and PnP-fbpconvnet method effectively reduced image noise and artifacts.

Figure 2-14 shows the low dose scan results (2mm Al). Due to the extremely low image quality of the FBP-2mm, post processing-based methods (BM3D-2mm and fbpconvnet-2mm) were ineffective in recovering the anatomy without adding artifacts. The TV-2mm blurred fine image structures. The PnP-fbpconvnet method achieved a better balance between retaining imaging details and minimizing image noise and artifacts. The superiority of the PnP-fbpconvnet over the fbpconvnet is attributed to the data consistency in model-based reconstruction.

Figure 2-15 shows the Noise power spectrum (NPS) of all reconstructed images. The BM3D method distorted the noise pattern compared with the original FBP image due to the non-linear denoising and lack of data consistency, potentially leading to image artifacts. The TV image removed high-frequency noise and image details. The noise patterns of fbpconvnet and PnP-fbpconvnet are similar to that of FBP.

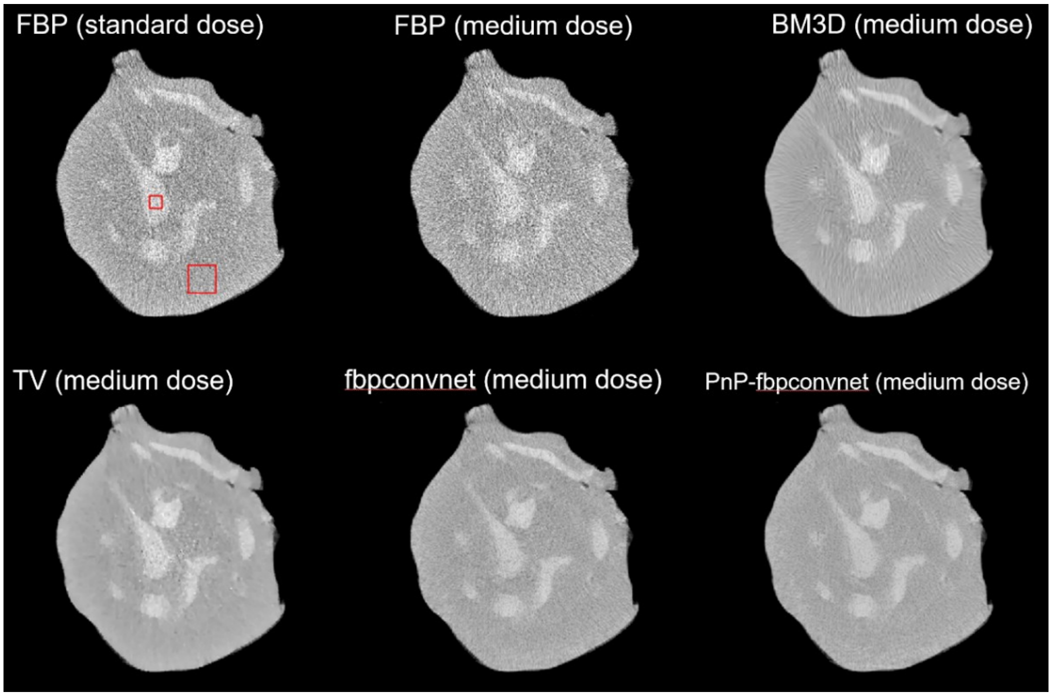


Figure 2-13 Reconstructed image of the breast phantom using FBP from standard dose scan, FBP, BM3D post-process, Total Variation, fbpconvnet, and PnP-fbpconvnet from medium dose scans. The conformal regions indicated by the red squares were used to evaluate the CNR.

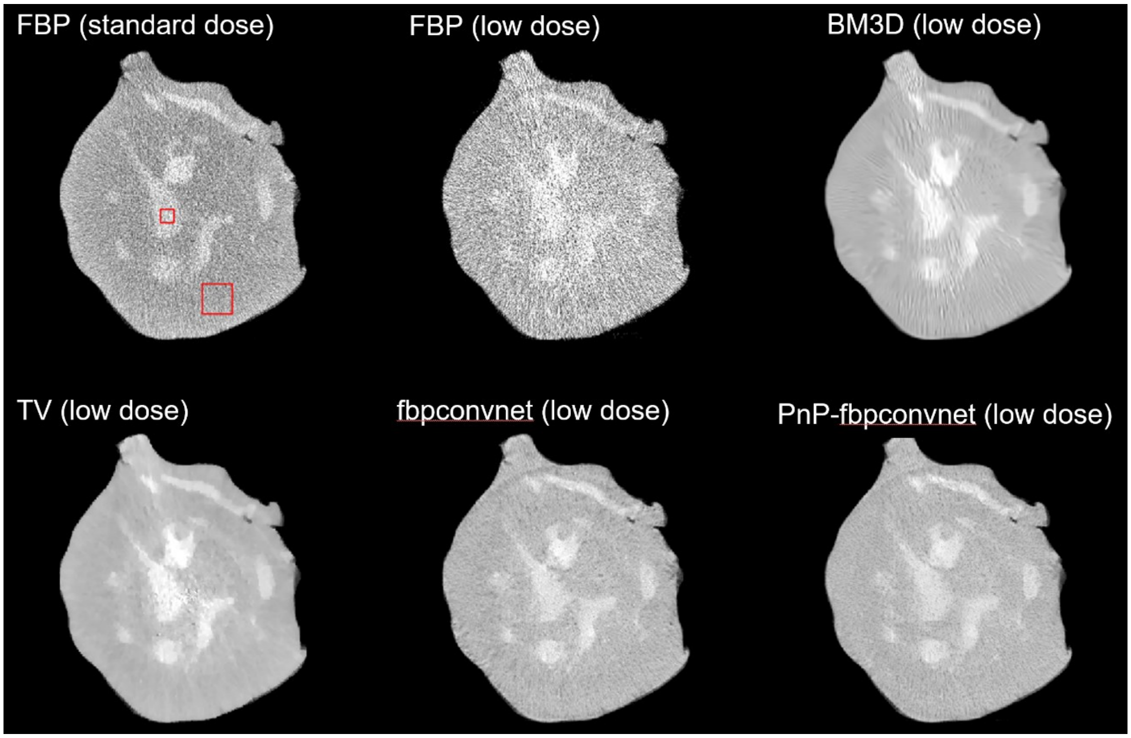


Figure 2-14 Reconstructed image of the breast phantom using FBP from standard dose scan, FBP, BM3D post-process, Total Variation, fbpconvnet, and PnP-fbpconvnet from low dose scans. The conformal regions indicated by the red squares were used to evaluate the CNR.

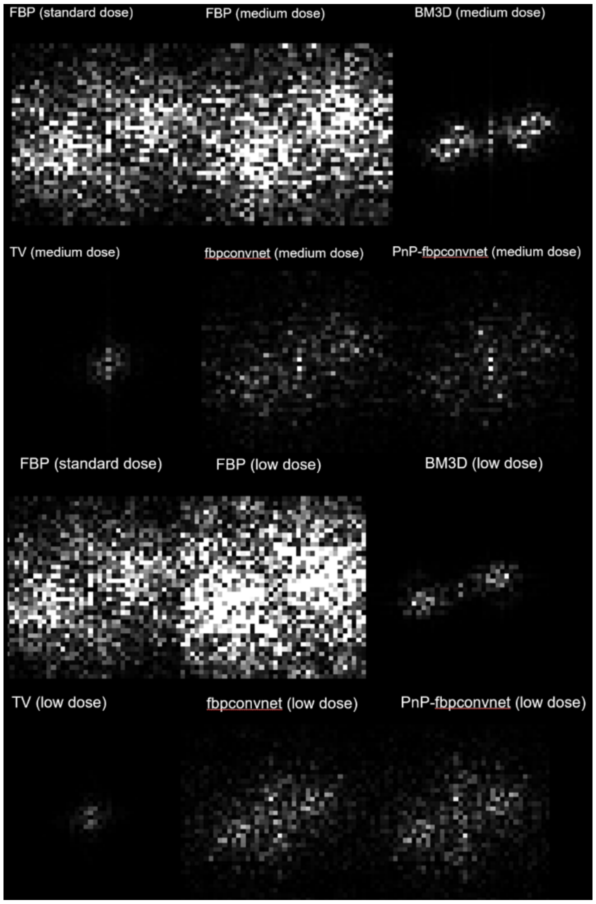


Figure 2-15 Noise power spectrum (NPS) of all reconstructed images.

Table 2-5 summarizes the contrast-to-noise ratio (CNR) evaluated based on two conformal regions, indicated by the red square on the reconstructed images shown in Figure 2-13 and Figure 2-14. BM3D and TV achieved the highest CNR but introduced image artifacts or blurred image. Compared with fbpconvnet, the PnP-fbpconvnet further improved CNR, showing the effectiveness of model-based reconstruction.

CNR	FBP	BM3D	TV	fbpconvnet	PnP-fbpconvnet
Low dose	1.01	5.62	4.48	3.54	3.81
Medium dose	0.92	2.59	2.54	2.39	3.29
Standard dose	1.12				

Table 2-5 The CNR of all methods evaluated on the conformal region indicated by the red square on Figure 2-13 and Figure 2-14.

2.3.4 Discussion

The PnP framework was initially proposed in [64] for model-based image restoration problems, which allows to plug in any off-the-shelf denoisers to replace a module in the alternating direction method of multipliers (ADMM) algorithm [28], and therefore combines iterative reconstruction with almost any state-of-the-art image denoisers without formulating the denoiser as an optimization problem. The flexible PnP framework has shown promising empirical results on many image restoration tasks [64–67].

In our previous study presented in section 2.2, we adapted the PnP-ADMM framework with BM3D for low dose CT iterative reconstruction. The PnP-ADMM-BM3D is shown to outperform the BM3D post-process method and conventional iterative reconstruction using TV regularization, showing the superiority of combining iterative reconstruction with state-of-the-art image denoisers. However, there are two major limitations with the previous PnP-ADMM-BM3D method. First, BM3D is ineffective at distinguishing image structures and artifacts and is subject to exaggerate the image artifacts on the input images. Second, the algorithm is computationally expensive since a linear equation involving the CT projection matrix is required to be solved at each iteration of ADMM. Despite the existence of an analytical solution to the linear equation, the solution can only be achieved with iterative algorithms (Conjugate Gradient algorithm) for unstructured system matrix such as the CT projection matrix. Unlike many image restoration problems where the multiplication with the system matrix and its transpose reduces to element-wise multiplication under Fourier Transform, the matrix multiplication is often the most time-consuming step in CT reconstruction and can be even more difficult for CBCT.

In this work, we adapted the PnP-PGD framework in[81] for LdCBBCT reconstruction and used DCNN as an image denoiser. Compared with PnP-ADMM, the PnP-PGD framework substantially reduced the number of matrix multiplication and improved computation efficiency. Compared with the PnP framework using BM3D as the denoiser, DCNN is a data-specific image denoiser and more effectively tackles the unique streaking artifacts and noises in LdCT.

2.3.5 Summary

In this study, a DCNN denoiser was plugged in the CTIR as an integrated reconstruction framework. By combining model-based optimization with deep learning, the PnP DCNN takes advantage of both the physical information in the CT forward model and the learned knowledge from the data-driven denoiser, showing improved SNR compared with both the DCNN based post-processing method and conventional IR methods. The flexible PnP algorithm is not limited to the denoiser used in this study, instead, it allows almost any image denoisers to be integrated in the CTIR framework.

3 USING DUAL-ENERGY CT TO GUIDE RADIATION THERAPY TREATMENT

3.1 Introduction

Despite the potential of using DECT for multi-material decomposition (MMD), the problem is ill-defined for decomposing more than two materials without additional assumption [19,20]. Existing DECT MMD approaches can be classified into three categories: projection-based method, integrated method, an image-based method. The projection-based method decomposes the two independent sinograms into two basis components by interpolating the lookup table and then performs separate image reconstruction [83]. The integrated method incorporates the DECT acquisition into the forward projection model and then reconstructs basis material images directly from the dual-energy CT projections [84]. The projection-based method and integrated method are robust to beam hardening artifacts [83,84], but there are several limitations. First, the projection-based method is limited to decomposing two basis materials. It is neither straightforward nor desirable to introduce additional constraints in the projection domain. Second, the projection-based method requires strict spatial and temporal consistency between the high and low energy acquisition. The condition is not met by many DECT systems using dual-source or single-source with fast kilovoltage-switching. Third, the integrated method is computationally expensive due to the repeated projection and back-projection during the reconstruction of basis image components. Fourth, the coupling between the reconstruction and image decomposition in the integrated method, especially with multiple basis materials, leads to decomposition results that are sensitive to parameter tuning.

Alternatively, image-based decomposition methods were investigated, which perform basis material decomposition on the high and low energy CT images. This method is applicable to all DECT systems, straightforward, and computationally tractable. Mendonca et al. [21] proposed an image-based MMD method that introduces two assumptions: mass and volume conservation, and that there are no more than three materials in each voxel. For each voxel, the method loops over a material triplet library, identifies the best-fit triplet, and decomposes the voxel into the three materials via direct matrix inversion. The method is termed Direct Inversion (DI), and it demonstrates the capability of decomposing CT into more than three basis materials. However, the decomposition accuracy is sensitive to the selection of material triplet library, which limits the flexibility of MMD, and the resultant images suffer from substantially amplified noise.

In this study, we proposed a DECT MMD method that decomposes the CT volume into multiple basis materials accurately while suppressing the decomposition image noise. Instead of rigidly confining each voxel to contain at most three basis materials, a non-convex sparsity term is introduced to penalize the number of materials that are simultaneously present in the same voxel.

3.2 Method

3.2.1 Formulation

The image domain DECT MMD problem is formulated as

$$\underset{x}{\text{minimize}} \qquad \frac{1}{2}\|Ax-b\|_2^2 + \lambda\,\mathrm{TV}(x) + \eta\sum_{k,i}|x_{ki}|^{\alpha}$$

$$\text{subject to} \qquad x \in S_P,$$

Equation 3-1

where the optimization variable x is the volume fraction (VF) matrix of dimension N_k by N_i. The voxels on the component image are indexed by i, and the decomposition materials are indexed by k. x_{ki} is the VF of the k^{th} material for the i^{th} voxel. A is the component-to-attenuation transformation matrix defined as

$$A = \begin{bmatrix} \mu_{1l} & \mu_{2l} & \cdots & \mu_{N_k l} \\ \mu_{1h} & \mu_{2h} & \cdots & \mu_{N_k h} \end{bmatrix},$$

Equation 3-2

where μ_{kl} and μ_{kh} are the LACs of the k^{th} material at low and high energies respectively. Matrix A transforms the VF into the LACs at high and low energies for each voxel. b is the attenuation image at high and low energies from measurements, defined as

$$b = \begin{bmatrix} b_{1l} & b_{2l} & \cdots & b_{N_i l} \\ b_{1h} & b_{2h} & \cdots & b_{N_i h} \end{bmatrix},$$

where b_{il} and b_{ih} denotes the LAC of the i^{th} voxel on the CT image at low and high energies, respectively. $TV(x)$ is the isotropic total variation regularization. S_P is the probability simplex set that enforces the mass and volume conservation rule on each voxel, such that each columns of the VF matrix x satisfies the sum-to-one and non-negative constraints

$$x \in S_P \quad \leftrightarrow \quad \begin{cases} x_{ki} \geq 0, & \forall i, k \\ \sum_k x_{ki} = 1, & \forall i. \end{cases}$$

The first term is the quadratic data fidelity term that minimizes the difference between the measured DECT image and the estimated DECT image calculated from the VF matrix x. The second term is the isotropic TV regularization term, applied on each component image to encourage image smoothness while preserving image edges. The last term is the penalty function in the form of $|\cdot|^\alpha$ $(0 \leq \alpha \leq 1)$, which promotes sparsity on the number of materials that simultaneously present in the same voxel. In this study, we specifically focus on four cases with $\alpha = 0, \frac{1}{2}, \frac{2}{3}, 1$. In the case of $\alpha = 1$, the sparsity term reduces to a constant under the mass and volume conservation constraint $(x \in S_P)$. In other cases, the penalty functions are non-convex, with the non-convexity increases as α goes to 0.

3.2.2 Algorithm

This study utilizes an accelerated primal-dual splitting approach with line search for both convex and non-convex problems [85,86], which solves the optimization problem of the form

$$\text{minimize} \;\; F(Kx) + G(x),$$

Equation 3-3

where G is convex, F possibly nonconvex, and K a linear operator.

The optimization problem in Equation 3-1 is formulated into the canonical form shown in Equation 3-3 by defining

$$K = \begin{bmatrix} A \\ D \\ I \end{bmatrix}, \qquad G(x) = I_{S_p}(x) = \begin{cases} \infty, & x \notin S_p \\ 0, & x \in S_p \end{cases}$$

$$F\left(\begin{bmatrix}\hat{z}_1\\ \hat{z}_2\\ \hat{z}_3\end{bmatrix}\right) = F_1(\hat{z}_1) + F_2(\hat{z}_2) + F_3(\hat{z}_3)$$

$$F_1(\hat{z}_1) = \frac{1}{2}\|\hat{z}_1 - b\|_2^2$$

$$F_2(\hat{z}_2) = \lambda\|\hat{z}_2\|_1$$

$$\left(F_3(\hat{z}_3)\right)_i = \eta|\hat{z}_{3i}|^{\alpha}.$$

Equation 3-4

$F(Kx)$ is equivalent to the objective function in equation (1). $G(x)$ is an indicator function that equals to 0 if $x \in S_p$ and infinity if $x \notin S_p$, which enforces the sum-to-one and non-negative constraints.

The accelerated primal-dual algorithm with line search[86] is presented in Algorithm 1, where the key steps are the evaluations of the proximal operator of function G and F.

The proximal operator of G reduces to the projection onto the probability simplex S_p, which can be efficiently solved by sorting and thresholding the input vector, as presented in Algorithm 2[87].

Algorithm 1: Accelerated primal-dual algorithm with line search

```
Initialization: $x_0 := 0 \in X,\ z_0 := 0 \in Z,\ t_0 > 0,\ \theta_0 := 1,\ \beta_0 := 1,\ \gamma > 0,\ r := 0.8$
For $\boldsymbol{k = 1, 2, \dots}$ do
        $x_k := \mathrm{Prox}_{t_{k-1},G}(x_{k-1} - t_{k-1}K^T z_{k-1})$
        $\boldsymbol{\beta_k := \beta_{k-1}(1 + \gamma t_{k-1})}$
        $\boldsymbol{t := t_{k-1}\sqrt{\frac{\beta_{k-1}}{\beta_k}(1 + \theta_{k-1})}}$
        Repeat
                $\theta_k := \frac{t}{t_{k-1}}$
                $\bar{x}_k := x_k + \theta_k(x_k - x_{k-1})$
                $s_k := \beta_k t$
                $\bar{z}_k := z_{k-1} + s_k K \bar{x}_k$
                $z_k := \bar{z}_k - s_k \mathrm{Prox}_{s_k^{-1},F}\left(\frac{\bar{z}_k}{s_k}\right)$
                $\mathrm{t_k := t}$
                Break if $\boldsymbol{\sqrt{\beta_k}t\left\|K^T(z_k - z_{k-1})\right\|_2 \le \|z_k - z_{k-1}\|_2}$
                $\boldsymbol{t := t_k * r}$
        End
End
```

Table 3-1 Accelerated primal-dual algorithm with line search.

Algorithm 2: Proximal operator evaluation of $\boldsymbol{G\ (\mathrm{Prox}_{tG}(x))}$

```
Input: $x \in R^{N_k} \times R^{N_i}$
For $\boldsymbol{i = 1, 2, \dots, N_i}$ do
        Sort $\{x_{1i}, x_{2i}, \cdots, x_{ki}, \cdots, x_{N_k i}\}$ into $x'$ such that $x'_{1i} \ge x'_{2i} \ge \cdots \ge x'_{N_k i}$
        Find $\boldsymbol{J_i := \max\left\{1 \le k \le K \middle| x'_{ki} + \frac{1}{k}\left(1 - \sum_{j=1}^{k} x'_{ji}\right)\right\}}$
        Define $\hat{x}_i := \frac{1}{J}\left(1 - \sum_{j=1}^{J} x'_{ji}\right)$
    For $k = 1,2, \dots, N_k$ do
            $\left(\mathrm{Prox}_{tG}(x)\right)_{ik} := \max\{x_{ik} + \hat{x}_i, 0\}$
Output: $\boldsymbol{\mathrm{Prox}_{tG}(x)}$
```

Table 3-2 Proximal operator evaluation of $\boldsymbol{G\ (\mathrm{Prox}_{tG}(x))}$

With the separable sum rule, evaluation of the proximal operator of F reduces to evaluating the proximal operator with respect to each variable

$$\mathrm{Prox}_{wF}\left(\begin{bmatrix}\hat{z}_1\\ \hat{z}_2\\ \hat{z}_3\end{bmatrix}\right) = \left(\begin{bmatrix}\mathrm{Prox}_{wF_1}(\hat{z}_1)\\ \mathrm{Prox}_{wF_2}(\hat{z}_2)\\ \mathrm{Prox}_{wF_3}(\hat{z}_3)\end{bmatrix}\right).$$

Following the definition of proximal operator in equation (4), the proximal operators of F_1 and F_2 are

$$\mathrm{Prox}_{wF_1}(\hat{z}_1) = \frac{\hat{z}_1 + wb}{w + 1},$$

$$\left(\mathrm{Prox}_{wF_2}(\hat{z}_2)\right)_{ik} = \begin{cases} \left(1 - \frac{w\lambda}{\|(\hat{z}_2)_{ik}\|_2}\right)(\hat{z}_2)_{ik}, & \|(\hat{z}_2)_{ik}\|_2 > w\lambda \\ 0, & \|(\hat{z}_2)_{ik}\|_2 \leq w\lambda. \end{cases}$$

Due to the separability, the proximal operator of F_3 reduces to pointwise proximal operator evaluation of $\eta|\cdot|^{\alpha}$

$$\left(\mathrm{Prox}_{wF_3}(\hat{z}_3)\right)_i = \underset{x}{\mathrm{argmin}} \left(\eta|x|^{\alpha} + \frac{1}{2w}(x - \hat{z}_{3i})^2\right)$$

Equation 3-5

Exact analytic solutions to Equation 3-5 exist for the scenarios considered in this study $\left(\alpha = 0, \frac{1}{2}, \frac{2}{3}, 1\right)$. Details can be found in the Appendix. For other values of α, the proximal operators could be evaluated numerically with an iterative approach such as the Newton's method [88].

3.2.3 Bone marrow decomposition using DECT MMD

Due to the distinct radiosensitivity of proliferating hematopoietic stem cells in active bone marrow, identifying and sparing active marrow may significantly improve radiotherapy treatment. ^{18}F-fluoro-l-deoxythymidine positron emission tomography (FLT-PET) differentiates active marrow from fatty marrow. However, FLT-PET is not FDA approved and cannot be used repeatedly for image-guided radiotherapy. Here we apply the MMD framework to perform bone marrow decomposition and compare it with PET and MRI images.

To separate calcium in the marrow, the DE ratio of calcium was included in the data fidelity in Equation 3-1:

$$\mathrm{A} = \begin{bmatrix} \mu_{hy} & \mu_{hr} & R_{cal} \\ \mu_{ly} & \mu_{lr} & 1 \end{bmatrix},$$

where R_{cal} is the DE ratio of calcium, defined as the slope of the high-low energy plot of calcium. μ_{hy} and μ_{ly} are the high and low energy LAC of yellow marrow, and μ_{hr} and μ_{lr} are the high and low energy LAC of red marrow.

3.2.4 Evaluation

The proposed framework with different sparsity parameters in the penalty functions was evaluated on a digital phantom, a Catphan©600 phantom, a Quantitative Imaging Phantom, and a pelvis patient and is compared with the direct inversion method [21]. The Quantitative Imaging Phantom data was acquired on a Siemens SOMATOM Force DECT. The Catphan and pelvis patient data were acquired on a Siemens SOMATOM Definition Flash. For the pelvis patient data, the mAs and CT Dose Index (CTDI) for the [low-energy, high-energy] acquisitions were [170mAs, 131mAs] and [4.92mGy, 3.66mGy] respectively. The low-energy and high-energy CT images were reconstructed using the standard FBP. The high-energy LAC and low-energy LAC of the basis material were computed as the average values of the high-energy LACs and low-energy LACs of the region of interest (ROIs), shown on the low-energy CT images for all cases.

For quantitative evaluation of the material decomposition accuracy, the mean and standard deviation (STD), electron density, as well as the VF were computed within each ROI on the decomposition component image.

The electron density ρ_i at voxel i is calculated by

$$\rho_i = \sum_{k=1}^{N_k} x_{ki}\rho(k),$$

where $\rho(k)$ is the electron density of the $\mathrm{k^{th}}$ material. The VF accuracy in a uniform ROI is defined as

$$\mathrm{VF} = \left(1 - \frac{\|\bar{x} - x_g\|_2}{\|x_g\|_2}\right) \times 100\%,$$

where $\bar{x}$ is the mean material component vector over all voxels within the uniform ROI, and x_g is the material component vector of the ground truth decomposition.

To quantify the amount of overlap between different material decompositions across the whole image, the Normalized Cross Correlation (NCC) coefficients at zero lag are evaluated for every pair of materials. The NCC coefficient at zero lag of material k_1 and k_2 is defined as

$$R_{k_1k_2} = \frac{\sum_{i=1}^{N_i} x_{k_1i}x_{k_2i}}{\sqrt{\sum_{i=1}^{N_i} x_{k_1i}^2}\sqrt{\sum_{i=1}^{N_i} x_{k_2i}^2}},$$

which equals to 0 if the two materials are completely separated and equals to 1 if they have identical distribution on the whole image. The NCC matrix is defined as a matrix with entries of $R_{k_1k_2}$. The diagonal elements of the NCC matrix are equal to 1 by definition, and the other elements are between 0 and 1, indicating the extent of material separation. If all materials are completely separated for all voxels, then all the off-diagonal elements should be 0. The diagonality D of the NCC matrix is computed using the Pearson correlation coefficient.

$$D = \frac{\left(\sum_{k_1,k_2} R_{k_1k_2}\right)\left(\sum_{k_1,k_2} k_1k_2R_{k_1k_2}\right) - \left(\sum_{k_1,k_2} k_1R_{k_1k_2}\right)\left(\sum_{k_1,k_2} k_2R_{k_1k_2}\right)}{\sqrt{\left(\sum_{k_1,k_2} R_{k_1k_2}\right)\left(\sum_{k_1,k_2} k_1^2R_{k_1k_2}\right) - \left(\sum_{k_1,k_2} k_1R_{k_1k_2}\right)^2}\sqrt{\left(\sum_{k_1,k_2} R_{k_1k_2}\right)\left(\sum_{k_1,k_2} k_2^2R_{k_1k_2}\right) - \left(\sum_{k_1,k_2} k_2R_{k_1k_2}\right)^2}}.$$

Equation 3-6

D equals to 1 for the diagonal matrix, -1 for the anti-diagonal matrix, and 0 for the uniform matrix. In the case of every voxel on the CT image being composed of only a single material, the diagonality of the NCC matrix is 1. On the other hand, in the case where every voxel contains an equal amount of all basis materials, the diagonality of the NCC matrix is 0. The diagonality D of the NCC matrix summarizes the decomposition separability over the whole image and across all materials.

For the bone marrow decomposition study, two subjects of the Wisconsin Miniature Swine (WMS)™ model were studied for their genetic proximity and similar anatomy to humans. For swine #1, CT, MRI, and PET images were obtained. For swine #2, 6 Gy of dose were given to the L4-L5 vertebrae of the swine. A total of 9 imaging sets in three imaging sessions were obtained, including before irradiation, one week after irradiation, and three weeks after irradiation. The CT images were acquired on a Siemens SOMATOM Definition Edge CT scanner. We evaluated the VF of active marrow within bone regions, including femur, ilium, sacrum ala, and vertebrae, using the proposed DECT decomposition method. The VF of active marrow evaluated from DECT was compared with that of the corresponding FLT-PET/MR and water-fat MRI scans.

3.3 Results

3.3.1 Digital Phantom

Figure 3-1 shows the low-energy and high-energy CT image of the digital phantom. This simple digital phantom is made up of four basis materials, including bone, iodine, water, and air, corresponding to the four ROIs indicated by the rectangular. Figure 3-2 shows the decomposition results on the digital phantom using the proposed framework with different α values and the classic direct inversion method. In the cases when $\alpha = 0, \frac{1}{2}, \frac{2}{3}$, the proposed framework distinctly separated all basis materials and with low noise, achieving a clear border between different materials. On the contrary, neither DI nor the proposed framework with $\alpha = 1$ achieved the desired sparsity or assigned the correct material to each ROI. The material separation capability of the proposed framework with $\alpha = 0, \frac{1}{2}, \frac{2}{3}$ is further confirmed with the NCC map in Figure 3-2, showing little or no crosstalk between two materials. With DI, the iodine component is mixed up with the bone and water. In the case of $\alpha = 1$, none of the components is separated.

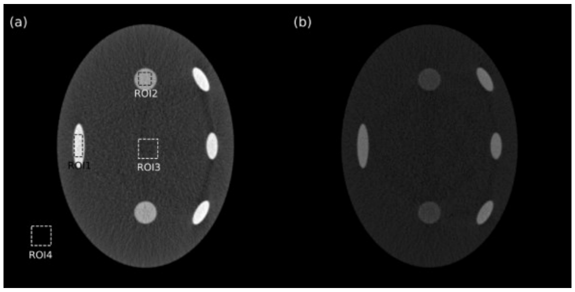


Figure 3-1 (a) The low energy: 75 kVp and (b) the high-energy: 140 kVp CT image of the digital phantom. The components of the ROIs are bone (ROI1), iodine (ROI2), water (ROI3), and air (ROI4). The displaying window is [0.01,0.065]mm^{-1}.

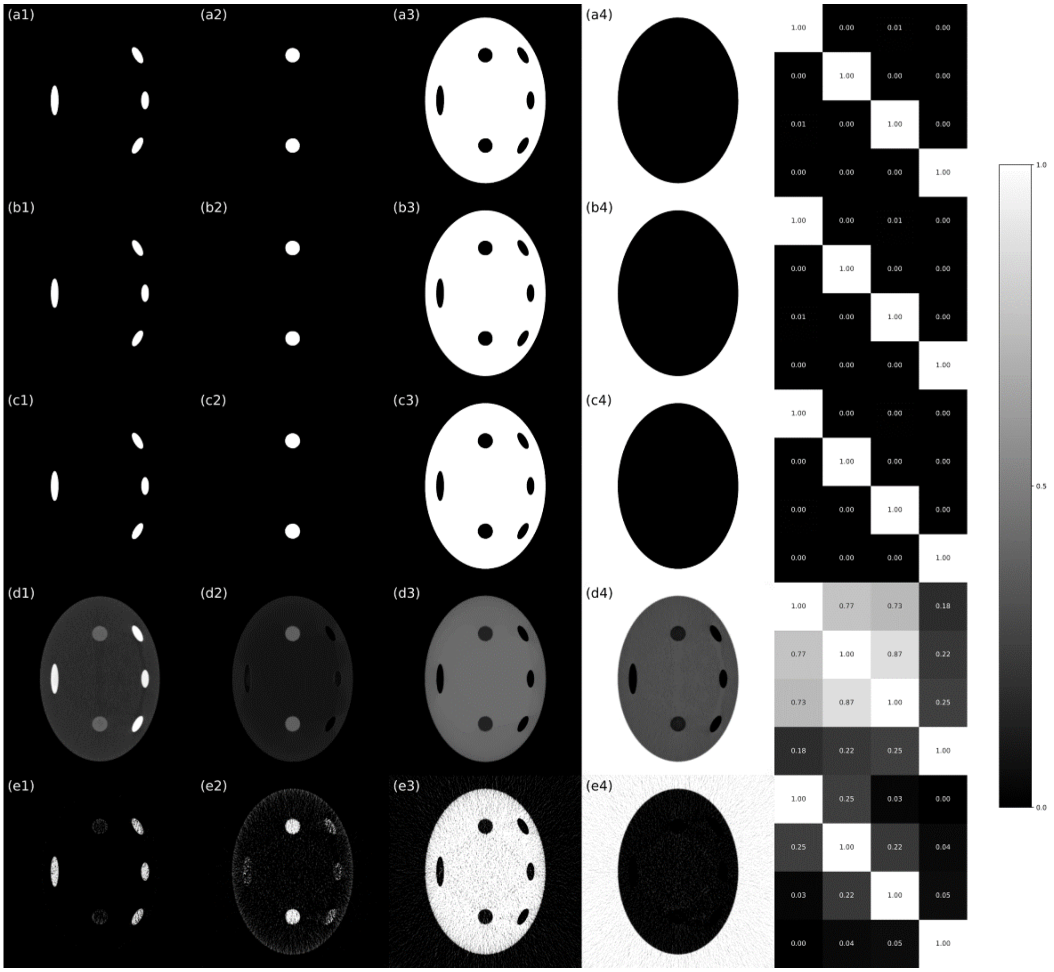


Figure 3-2 Decomposition component images of (1) bone (2) iodine (3) water and (4) air, decomposed using the proposed framework when (a) $\alpha = 0$, (b) $\alpha = \frac{1}{2}$, (c) $\alpha = \frac{2}{3}$, (d) $\alpha = 1$, and (e) the Direct Inversion method. The last column is the 4 by 4 NCC map of the decomposition in the same row, with each square showing the corresponding entries in the NCC matrix, where the basis materials are bone, iodine, water, and air, from top to bottom and from left to right.

3.3.2 Catphan©600 phantom

Figure 3-3 shows the low- and high-energy CT images of the Catphan©600 phantom with contrast rods, which are made of six basis materials including Teflon, Delrin, Iodine solution of 10 mg/ml, PMP, Inner soft tissue, and Air, corresponding to the six ROIs on the low energy CT image (Figure 3-3(a)). The materials in the labeled rods on the high energy CT (Figure 3-3(b)) are Teflon, Delrin, Iodine solution of 10 mg/ml, Polystyrene, low-density

Polyethylene (LDPE), Polymethyl pentene (PMP), and Iodine solution of 5 mg/ml, respectively. The VF accuracy was evaluated on the six ROIs, and the electron density was evaluated on the seven contrast rods. Figure 3-4 shows the decomposition image and the NCC map. The proposed framework with $\alpha = 0, \frac{1}{2}, \frac{2}{3}$ successfully separated the phantom into the six basis components with minimal crosstalk between different materials. With the classical DI method, the off-diagonal elements of the NCC matrix are up to 0.36, showing that the two corresponding basis materials, Delrin and PMP, are not well-separated. The proposed method with $\alpha = 1$ is unable to achieve material separation at all. Table 3-3 shows the evaluated electron densities for the seven contrast rods. The mean square error of the electron density was reduced by 72.6% for the proposed framework with $\alpha = 0, 1/2, 2/3$ compared with DI.

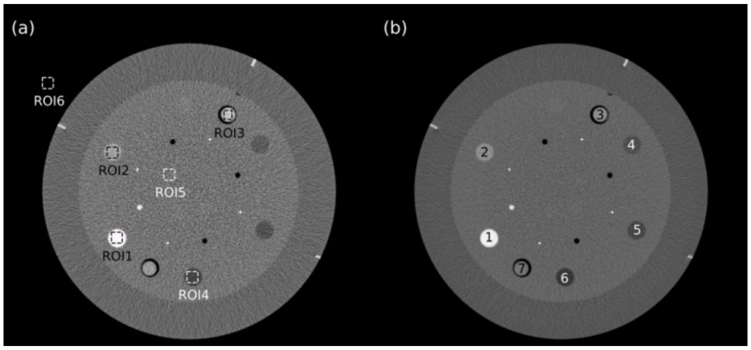
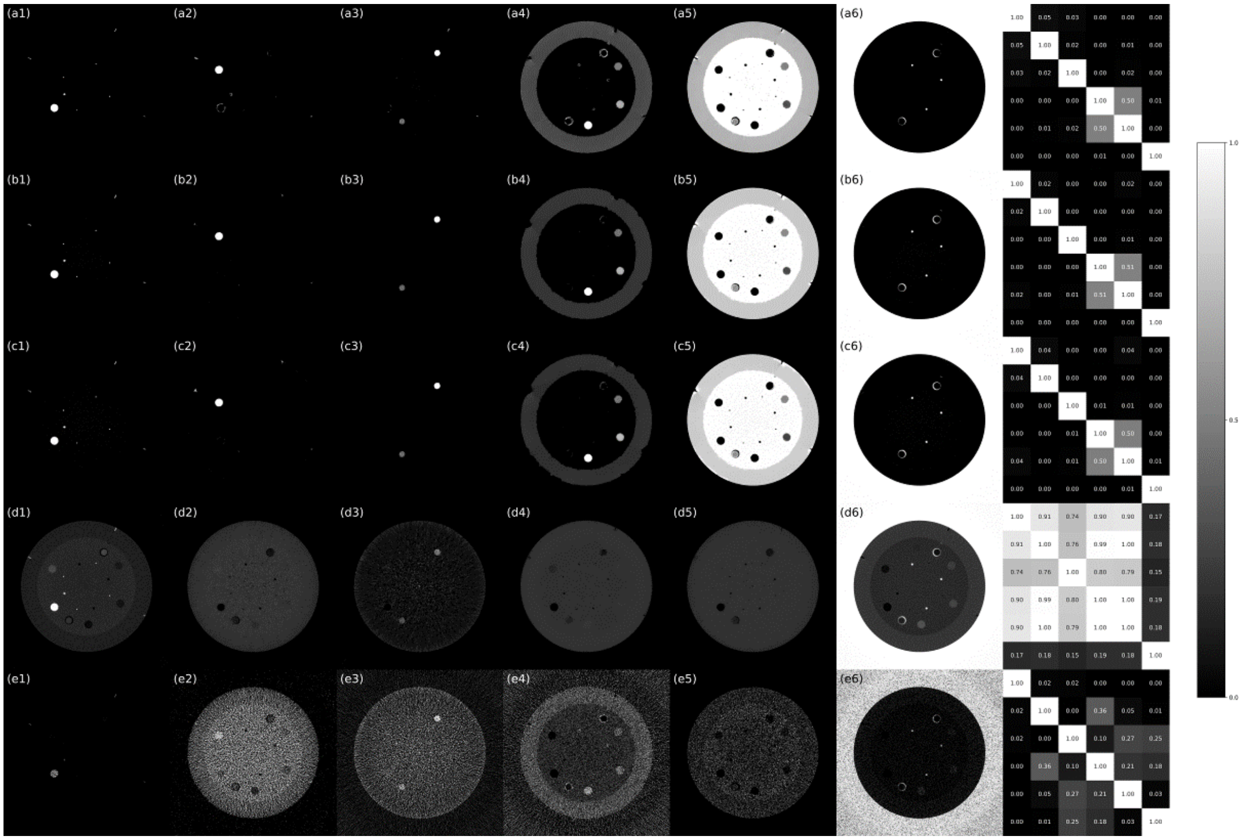


Figure 3-3 (a) The low energy: 75 kVp and (b) the high-energy: 125 kVp CT image of the Catphan©600 phantom. The material components of the ROIs on the low energy CT image are Teflon (ROI1), Delrin (ROI2), Iodine of 10 mg/ml (ROI3), PMP (ROI4), Inner soft tissue (ROI5) and Air (ROI6). The labeled contrast rods on the high energy CT image are composed of (1) Teflon, (2) Delrin, (3) Iodine solution of 10 mg/ml, (4) Polystyrene, (5) low-density Polyethylene (LDPE), (6) Polymethyl pentene (PMP), (7) Iodine solution of 5 mg/ml. The displaying window is [0.01,0.04]mm^{-1}.



1.00	0.05	0.03	0.00	0.00	0.00
0.05	1.00	0.02	0.00	0.01	0.00
0.03	0.02	1.00	0.00	0.02	0.00
0.00	0.00	0.00	1.00	0.50	0.01
0.00	0.01	0.02	0.50	1.00	0.00
0.00	0.00	0.00	0.01	0.00	1.00

1.00	0.02	0.00	0.00	0.02	0.00
0.02	1.00	0.00	0.00	0.00	0.00
0.00	0.00	1.00	0.00	0.01	0.00
0.00	0.00	0.00	1.00	0.51	0.00
0.02	0.00	0.01	0.51	1.00	0.00
0.00	0.00	0.00	0.00	0.00	1.00

1.00	0.04	0.00	0.00	0.04	0.00
0.04	1.00	0.00	0.00	0.00	0.00
0.00	0.00	1.00	0.01	0.01	0.00
0.00	0.00	0.01	1.00	0.50	0.00
0.04	0.00	0.01	0.50	1.00	0.01
0.00	0.00	0.00	0.00	0.01	1.00

1.00	0.91	0.74	0.90	0.90	0.17
0.91	1.00	0.76	0.99	1.00	0.18
0.74	0.76	1.00	0.80	0.79	0.15
0.90	0.99	0.80	1.00	1.00	0.19
0.90	1.00	0.79	1.00	1.00	0.18
0.17	0.18	0.15	0.19	0.18	1.00

1.00	0.02	0.02	0.00	0.00	0.00
0.02	1.00	0.00	0.36	0.05	0.01
0.02	0.00	1.00	0.10	0.27	0.25
0.00	0.36	0.10	1.00	0.21	0.18
0.00	0.05	0.27	0.21	1.00	0.03
0.00	0.01	0.25	0.18	0.03	1.00

Figure 3-4 Decomposition component images of (1) Teflon, (2) Delrin, (3) Iodine solution of 10 mg/ml, (4) PMP, (5) Inner soft tissue (LDPE), (6) Air, decomposed using the proposed framework when (a) $\alpha = 0$, (b) $\alpha = \frac{1}{2}$, (c) $\alpha = \frac{2}{3}$, (d) $\alpha = 1$, and (e) the Direct Inversion method. The last column is the 6 by 6 NCC map of the decomposition in the same row, with each square showing the corresponding entries in the NCC matrix, where the basis materials are Teflon, Delrin, Iodine solution of 10 mg/ml, PMP, LDPE, and air, from top to bottom and from left to right.

Method		Rod 1 Teflon	Rod 2 Delrin	Rod 3 Iodine (10 mg/ml)	Rod 4 Polystyrene	Rod 5 LDPE	Rod 6 PMP	Rod 7 Iodine (5 mg/ml)	Average NMSE
Ground Truth	ρ_e	6.240	4.525	3.368	3.400	3.155	2.851	3.356	\
$\alpha = 0$	ρ_e NMSE	6.240 0.00%	4.525 0.00%	3.368 0.00%	3.104 8.70%	3.008 4.67%	2.851 0.00%	3.354 0.06%	1.92%
$\alpha = 1/2$	ρ_e NMSE	6.218 0.35%	4.514 0.24%	3.357 0.33%	3.119 8.26%	2.997 5.01%	2.854 0.12%	3.352 0.13%	2.06%
$\alpha = 2/3$	ρ_e NMSE	6.211 0.46%	4.507 0.41%	3.346 0.66%	3.113 8.43%	2.984 5.43%	2.856 0.16%	3.346 0.30%	2.27%
$\alpha = 1$	ρ_e NMSE	6.169 1.14%	4.203 7.11%	4.011 19.11%	2.961 12.91%	2.770 12.22%	2.509 12.01%	3.578 6.63%	10.16%
Direct Inversion	ρ_e NMSE	4.847 22.33%	4.174 7.76%	3.699 9.83%	3.251 4.38%	3.134 0.67%	2.887 1.28%	3.523 4.96%	7.32%

Table 3-3 The electron densities measured on the Catphan contrast rods labeled on Figure 3-3(b). The RMSE is evaluated for each method as the mean square error of the seven rods.

3.3.3 Quantitative Imaging phantom

Figure 3-5 show the DECT image for the Quantitative Imaging Phantom, which consists of 12 basis materials including 2 mg/ml iodine solution (ROI1), 5 mg/ml iodine solution (ROI2), 10 mg/ml iodine solution (ROI3), 15 mg/ml iodine solution (PMP) (ROI4), 50 mg/ml calcium solution (ROI5), 100 mg/ml calcium solution (ROI6), 300 mg/ml calcium solution (ROI7), HE blood 70 (ROI8), HE blood 100 (ROI9, adipose (ROI10), water (ROI11), and brain (ROI12). In this phantom study, to avoid being trapped in undesired local minima with the increased number of basic materials, the algorithm is initialized by setting VF=1 for the basis material that is closest to each pixel. Figure 3-6 shows the decomposition images with the corresponding NCC map. The proposed framework with $\alpha = 0, \frac{1}{2}, \frac{2}{3}$ decomposed the DECT images into 12 different basis components plus air and improved the VF accuracy from 51% using DI method to 100%. The non-diagonal entries of the NCC map are close to 0 for the proposed framework with $\alpha = 0, \frac{1}{2}, \frac{2}{3}$, showing clean separation of the basic materials. The off-diagonal entries are close to 1 for the $\alpha = 1$ scenario. The classical DI method is unable to separate similar materials. For example, the HE blood 100 (ROI9), adipose (ROI10), and water (ROI11) have similar LAC values, resulting in substantial non-zero NCC elements $R_{k_1 k_2}$ using DI.

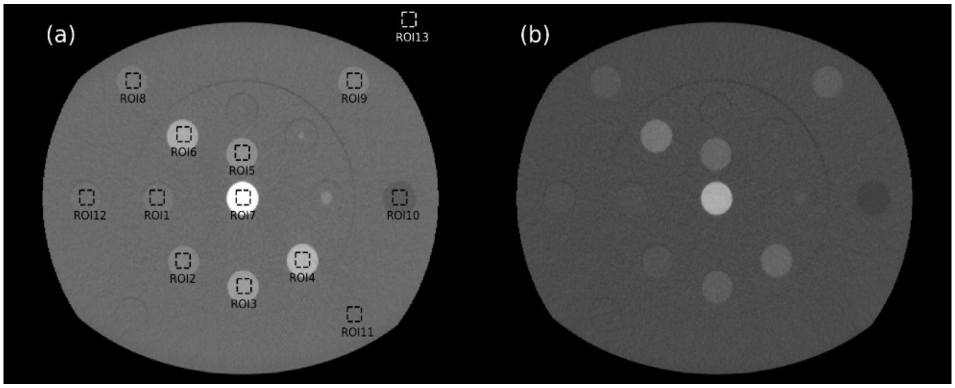


Figure 3-5 (a) The low energy: 100 kVp and (b) the high-energy: 150 kVp CT image of the Quantitative Imaging phantom. The material components of the ROIs are 2 mg/ml iodine solution (ROI1), 5 mg/ml iodine solution (ROI2), 10 mg/ml iodine solution (ROI3), 15 mg/ml iodine solution (PMP) (ROI4), 50 mg/ml calcium solution (ROI5), 100 mg/ml calcium solution (ROI6), 300 mg/ml calcium solution (ROI7), HE blood 70 (ROI8), HE blood 100 (ROI9, adipose (ROI10), water (ROI11), brain (ROI12), air (ROI13).

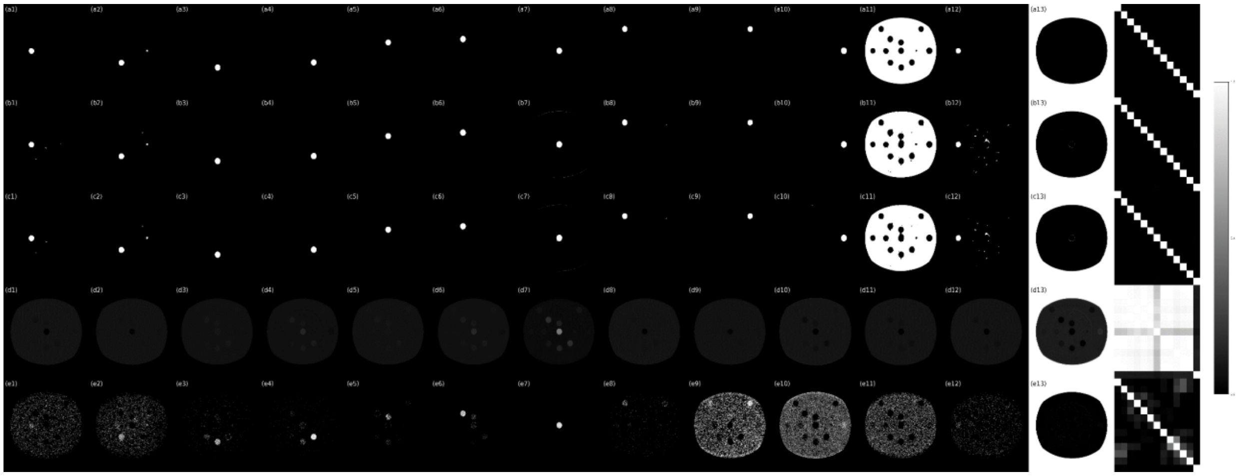


Figure 3-6 Decomposition component images of (1) 2 mg/ml iodine solution, (2) 5 mg/ml iodine solution (ROI2), (3) 10 mg/ml iodine solution (ROI3), (4) 15 mg/ml iodine solution (PMP) (ROI4), (5) 50 mg/ml calcium solution (ROI5), (6) 100 mg/ml calcium solution (ROI6), (7) 300 mg/ml calcium solution (ROI7), (8) HE blood 70 (ROI8), (9) HE blood 100 (ROI9), (10) adipose (ROI10), (11) water (ROI11), (12) brain (ROI12), (13) air (ROI13), decomposed using the proposed framework when (a) $\alpha = 0$, (b) $\alpha = \frac{1}{2}$, (c) $\alpha = \frac{2}{3}$, (d) $\alpha = 1$, and (e) the Direct Inversion method. The last column is the NCC map of the corresponding decomposition, with each square showing one NCC matrix element.

3.3.4 Pelvis patient

Figure 3-7 and Figure 3-8 show the DECT image and the decomposition images with the corresponding NCC map respectively for the pelvis patient. The proposed framework with

$\alpha = 0, \frac{1}{2}, \frac{2}{3}$ decomposed the DECT images into bone (ROI1), iodine (ROI2), muscle (ROI3), fat (ROI4), and air, and achieved an NCC matrix with non-diagonal coefficients close to 0.

Table 3-4 summarizes the VF accuracy and diagonality of the NCC matrix for all datasets. The proposed method with $\alpha = 0, \frac{1}{2}, \frac{2}{3}$ achieves a significantly higher VF accuracy and diagonality than the comparison methods and approaching nearly perfect material decomposition. Across all datasets, the proposed method improved the average VF accuracy from 61.2% to 99.9% and increased the diagonality of the NCC matrix from 0.73 to 0.96.

Table 3-5 presents the runtime and the hyperparameters used for all cases. Despite that our algorithm used MATLAB built-in GPU computing tools for acceleration, the proposed framework takes 12 minutes on average, which is slower than the direct inversion method, requiring only 3 minutes on CPU. The hyperparameters were tuned case by case to achieve visually desired sparseness and smoothness. The sparseness and the smoothness can be promoted by increasing η and λ respectively. γ and t are related to the step sizes in the algorithms, which were tuned in a trial-and-error way for faster and more stable convergence.

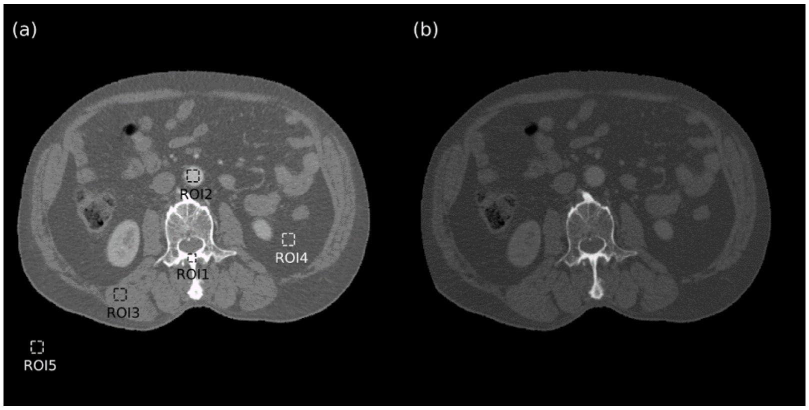


Figure 3-7 (a) The low energy: 100 kVp and (b) the high-energy: 140 kVp CT image of the pelvis patient. The material components of the ROIs are bone (ROI1), iodine (ROI2), muscle (ROI3), fat (ROI4), and air (ROI5). The displaying window is [0.01,0.04]mm^{-1}.

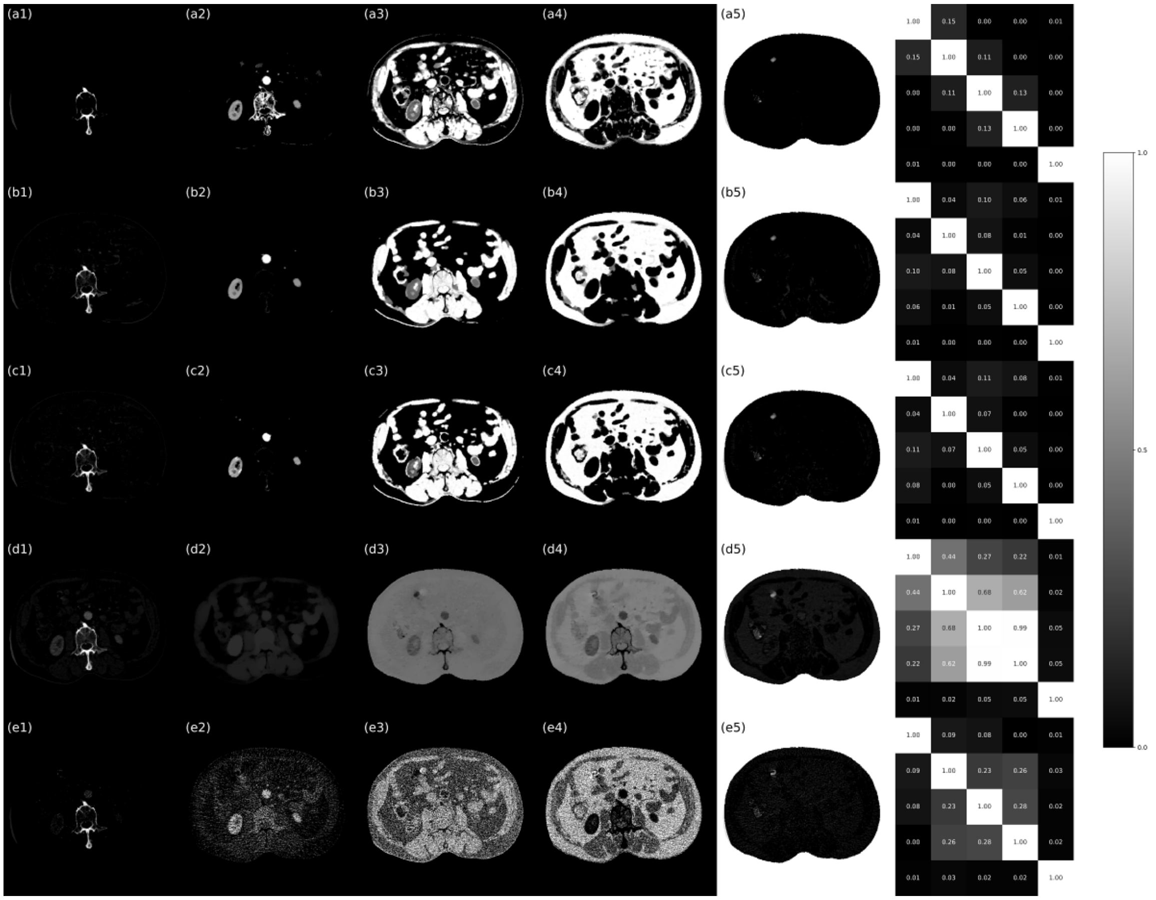


Figure 3-8 Decomposition component images of (1) bone (2) iodine (3) muscle (4) fat and (5) air, decomposed using the proposed framework when (a) $\alpha = 0$, (b) $\alpha = \frac{1}{2}$, (c) $\alpha = \frac{2}{3}$, (d) $\alpha = 1$, and (e) the Direct Inversion method. The last column is the 5 by 5 NCC map of the decomposition in the same row, with each square showing the corresponding entries in the NCC matrix, where the basis materials are bone, iodine, muscle, fat, and air, from top to bottom and from left to right.

		Digital Phantom	Catphan	Quantitative Imaging phantom	pelvis patient
VF accuracy	$\alpha = 0$	100.0%	99.9%	100%	NA
	$\alpha = 1/2$	100.0%	99.6%	100%	
	$\alpha = 2/3$	100.0%	99.3%	100%	
	$\alpha = 1$	62.1%	45.3%	15%	
	Direct	88.4%	44.1%	51%	
Diagonality	$\alpha = 0$	1.00	0.95	1.00	0.95
	$\alpha = 1/2$	0.99	0.94	0.99	0.89
	$\alpha = 2/3$	0.99	0.93	0.99	0.87
	$\alpha = 1$	0.33	0.14	0.03	0.53
	Direct	0.87	0.72	0.54	0.8

Table 3-4 The Volume Fraction accuracy and Diagonality of the NCC matrix for all datasets.

		Digital Phantom	**Catphan**	**Quantitative Imaging phantom**	**pelvis patient**
η	$\alpha = 0$	0.03	0.006	0.0009	0.003
	$\alpha = 1/2$	0.05	0.002	0.0009	0.001
	$\alpha = 2/3$	0.1	0.003	0.0009	0.002
	$\alpha = 1$	0.5	0	0.1	0
λ	$\alpha = 0$	0.05	0.002	0.02	0.0007
	$\alpha = 1/2$	0.25	0.003	0.005	0.0012
	$\alpha = 2/3$	0.05	0.0025	0.003	0.0005
	$\alpha = 1$	0.01	0.0005	0	0.0005
γ	$\alpha = 0$	0.01	0.0001	1.00E-05	0.0001
	$\alpha = 1/2$	0.01	0.0001	1.00E-05	0.0001
	$\alpha = 2/3$	0.001	0.001	0.001	1.00E-05
	$\alpha = 1$	0.1	0.001	1	0.001
t	$\alpha = 0$	100	1	1	1
	$\alpha = 1/2$	100	1	1	1
	$\alpha = 2/3$	100	100	0.5	100
	$\alpha = 1$	100	100	0.01	100
Runtime (s)	$\alpha = 0$	63.4	791.5	861.7	738.8
	$\alpha = 1/2$	71.2	881.4	489.1	1640.1
	$\alpha = 2/3$	81.5	1059.4	3542.8	962.1
	$\alpha = 1$	81.3	95.1	323.9	294.7
	Direct Inversion	36.2	133.6	490.1	30.9

Table 3-5 The runtime and the hyperparameters used for all cases.

Figure 3-9 shows the convergence plots for different α values on the Catphan and the pelvis patient case. The objective values for different cases were not comparable, since different hyperparameter values were used in different cases. However, it is worth to note that the plots show different converging patterns. For the convex case $\alpha = 1$, the objective goes down nicely with a convex-shaped convergence curve, showing stable and robust convergence. For the non-convex case $\alpha = 0$, the curve goes down with sudden changes. For the α values in between, the $\alpha = \frac{2}{3}$ curve patterns are more similar to that of the $\alpha = 1$ curves, while the $\alpha = \frac{1}{2}$ curves are more irregular and bumpier.

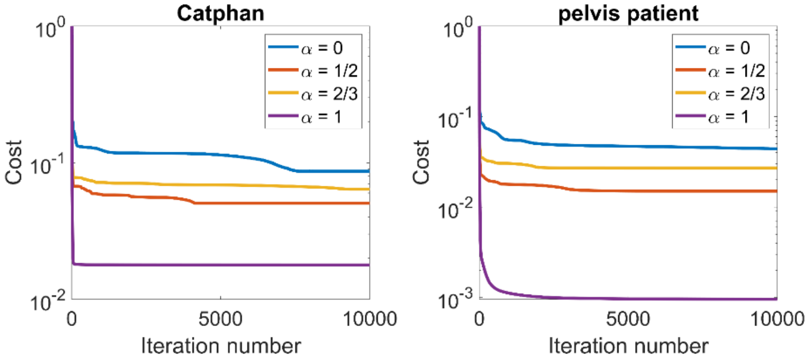


Figure 3-9 Convergence plots for different α values on the Catphan and the pelvis patient case.

3.3.5 Bone marrow decomposition

Figure 3-10 and Figure 3-11 show the volume fraction of active marrow evaluated for various bones using MRI, CT, and PET images. The bone marrow was decomposed into active marrow, fatty marrow, and calcium. The active marrow component computed from DECT accord with that of FLT-PET (ground truth). The volume fraction of active marrow evaluated from [MRI, DECT, PET] are [0.20, 0.04, 0.05], [0.16,0.14,0.17], [0.51, 0.78, 0.54], and [0.51, 0.85, 0.36] for femur, ilium, sacrum ala, and vertebrae, respectively. Vertebrae and sacrum ala have the highest concentration of active marrow, shown on MRI, DECT and PET. Femur has the lowest concentration of active marrow, shown on the DECT and PET. The correlation is 0.90 between PET and MR, and 0.92 between PET and DECT.

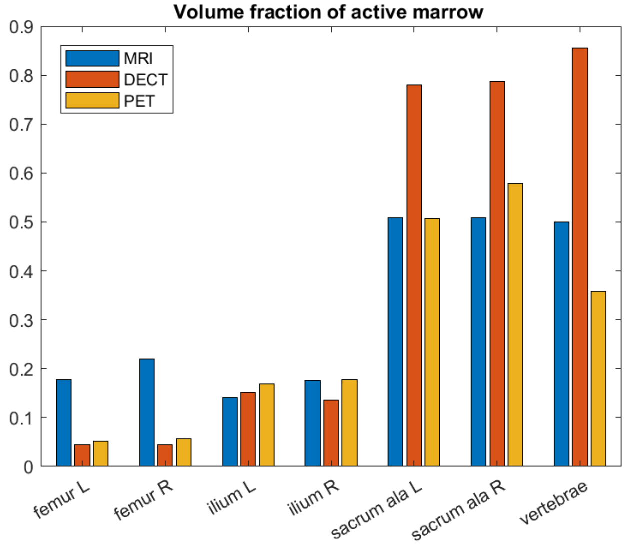


Figure 3-10 Volume fraction of active marrow evaluated for various bones using MRI, CT, and PET images.

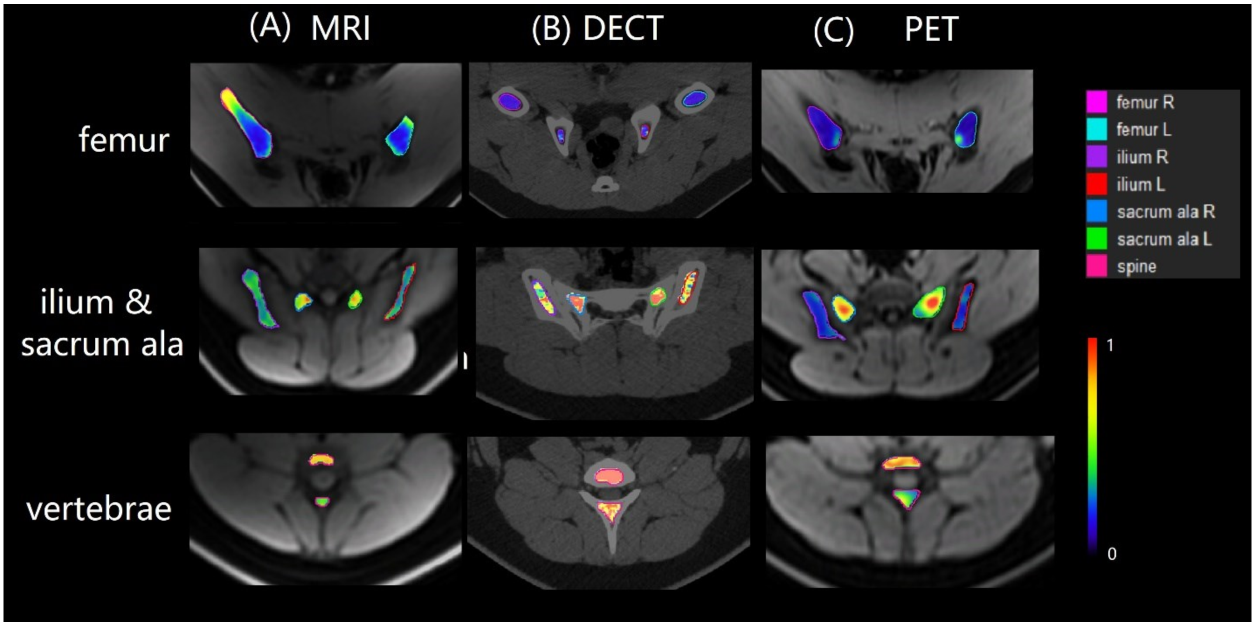


Figure 3-11 Color-wash images of active marrow volume fraction computed from (A) water-fat MR, (B) DECT decomposition, (C) PET. The color-wash images were fused to (A) water MRI, (B) high-energy CT, (C) MRI acquired from PET/MR. The PET was normalized by the maximal PET signal within the bone marrow. The colors of structure contours are shown in the legend.

Figure 3-12 shows the PET SUV and the active marrow volume fraction of MRI and DECT for three imaging sessions including before irradiation, one week after irradiation, and three weeks after irradiation. For the bone marrow regions received radiation including ilium, sacrum ala, and vertebrae, the volume fraction of active marrow and the PET SUV signals decreases after radiation. For the unirradiated bone marrow within the femur, the active marrow volume fraction remains unchanged.

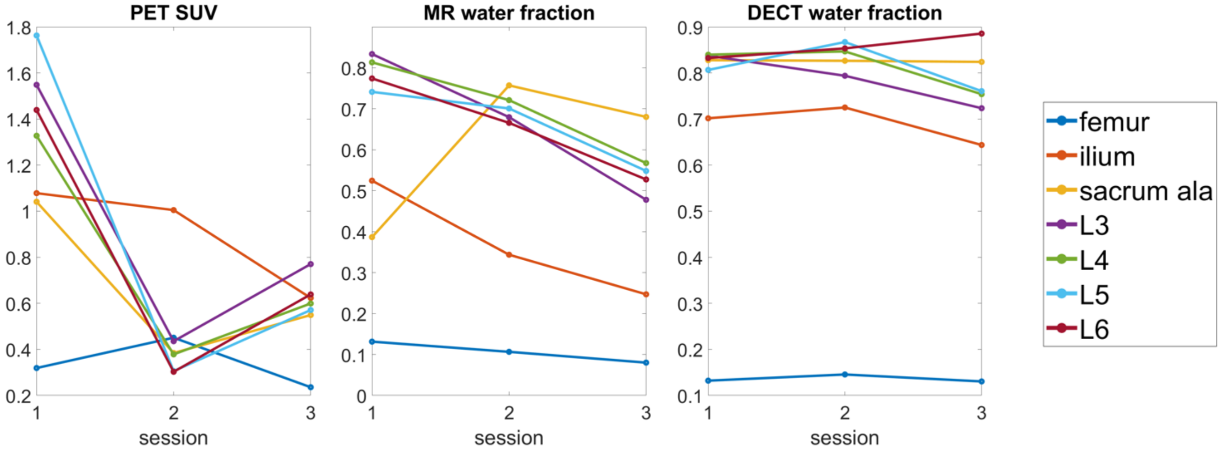


Figure 3-12 The PET SUV and the active marrow volume fraction of MRI and DECT.

3.4 Discussion

The standard DI method for DECT based MMD imposes sparsity constraint by enforcing a hard ceiling of three on the number of materials in each voxel and then solves the basis material components by direct matrix inversion. However, the three-material constraint is arbitrary and unrealistically rigid. Moreover, the direct matrix inversion inevitably amplifies the image noise as shown in previous publications [21] and the current study. Our proposed DECT MMD framework utilizes a TV regularization that regulates the decomposition image noise and uses a sparsity regularization to penalize the number of materials that are simultaneously present at the same voxel. The soft sparsity regularization allows the number of basis materials to vary across different voxels.

The sparsity term is in the form of $|x|^{\alpha}$ $(0 \leq \alpha \leq 1)$, where $\alpha = 0, \frac{1}{2}, \frac{2}{3}, 1$ were specifically studied. The sparsity term reduces to the L1 norm of the x with $\alpha = 1$, and further reduces to a constant under the mass and volume constraint. Therefore, the sparsity term with $\alpha = 1$ does not promote material sparsity despite its desirable mathematical properties of being convex. When $\alpha = 0, \frac{1}{2}, \frac{2}{3}$, the corresponding sparsity term has a closed-form proximal operator, which could be difficult to evaluate for other values of α. The differences between $\alpha = 0$, $\alpha = \frac{1}{2}$, and $\alpha = \frac{2}{3}$ are subtle with respect to the decomposition results. The material penalty term with $\alpha = 0$ is also referred to as the counting norm and a mathematically rigorous description of the material sparsity. This is reflected in the best quantitative performance achieved with this norm. However, due to its extreme non-convexity, it is more challenging to tune the parameters for convergence to an acceptable local minimum. The sparsity terms with $\alpha = \frac{1}{2}$ and $\alpha = \frac{2}{3}$ are better behaving non-convex functions, showing more robust performance to optimization parameter selection.

The non-convex optimization problem in this study was solved using an accelerated primal-dual algorithm with line search proposed in [86], which reduces to the Primal-Dual Hybrid Gradient (PDHG) Method [29,89] when applied to convex optimization problems. As discussed in Section 1.5, FISTA is a faster algorithm that applies to most of the studies in this thesis. We will investigate FISTA for the non-convex optimization problem in the future.

3.5 Conclusions

The proposed method accurately decomposed the dual energy CT phantom and patient images into up to 12 basis materials, markedly reduced crosstalk among materials, suppressed decomposition image noise, and retained image spatial resolution. The proposed method using non-convex material sparsity penalty outperforms convex penalty and the standard direct inversion method. When applied to bone marrow decomposition, DECT identifies active bone marrow that is consistent with the FLT-PET and water-fat MRI.

4 NEW DEGREES OF FREEDOM IN VMAT

4.1 An integrated and flexible framework of Volumetric Modulated Arc Therapy

4.1.1 Introduction

Rotational intensity modulated radiotherapy was initially described by Mackie et al. (Mackie, 1993) for the proprietary Helical TomoTherapy hardware platform. For the more commonly available C-arm gantry systems, Yu[23] introduced the intensity-modulated arc therapy (IMAT) concept, which uses multiple arcs to substitute the conventional static beam intensity modulated radiotherapy (IMRT) with the promise of improving delivery efficiency and robustness using simpler Multi-leaf Collimator (MLC) segments. Several modifications and improvements were subsequently made on the original IMAT modulation to improve both the dosimetric quality and reduce the number of arcs [90–93].

Compared with the static beam IMRT, Volumetric Modulated Arc Therapy (VMAT) optimization is more challenging due to its large problem size and more complex constraints. Typically, 180 or more beams are included in the VMAT delivery compared with fewer than ten beams used in a typical IMRT plan. More importantly, the gantry rotation and leaf motion is coupled: for efficient VMAT delivery, the MLC leaf movements between adjacent gantry angles cannot exceed the product of the maximal leaf speed and the gantry travel time, which is short to maintain smooth and efficient gantry rotation. The progressive sampling optimization (PSO) method for VMAT proposed by Otto[22] successfully addressed the mechanical constraint problem and kept the computational complexity manageable. The PSO method solves VMAT optimization using coarsely sampled control points in the initial stage, and then progressively inserts new beams until reaching the final sampling resolution, e.g., 2 degrees. The PSO method ensures a deliverable VMAT plan that meets MLC mechanical constraints and is adopted in many current commercial planning systems. However, the heuristic PSO method has several intrinsic limitations. First, the interpolation during the progressive sampling procedure routinely leads to MLC aperture forming contention issues. Large, unsynchronized MLC motion between coarsely sampled control points results in unnecessary dose deposition to healthy tissue[94]. Second, the optimization result is heavily influenced by the parameters at the initial coarse sampling stage. The algorithm is easily trapped in undesired local minima with suboptimal parameters at the initial stage. Subsequently, despite the original premise of single-arc VMAT, two or more arcs are typically required in practice for acceptable dosimetric quality. Third, the optimization results are not only dependent on the parameters themselves but also on the entire parameter tuning history. It is, therefore, difficult to reproduce a plan even if the final set of optimization parameters is provided. Last but not least, the heuristic algorithm is designed specifically for the current VMAT delivery method and machine hardware.

In this study, we develop an integrated VMAT optimization framework with global sampling that not only improves dosimetry but also is flexible for a variety of advances in radiotherapy treatment and delivery.

4.1.2 Methods

4.1.2.1 <u>Optimization Formulation</u>

The proposed comprehensive VMAT (comVMAT) optimization formulation takes the following form:

$$\underset{\{f_\theta\,, c_\theta, \Phi_\theta\}_{\theta=0}^{n}}{argmin} \quad \frac{1}{2}\left\| W\left(\sum_{\theta}(A_\theta f_\theta) - d_0\right)\right\|_2^2 + \sum_{\theta}\underbrace{\left(\lambda_1\|D_{1\theta} f_\theta\|_1 + \lambda_2\|D_{2\theta} f_\theta\|_1\right)}_{anisotropic\ TV\ term\ on\ f}$$

$$+\sum_{\theta}\sum_{x,y}\left(\underbrace{\frac{\gamma}{2}\left[\left(f_{\theta xy} - c_\theta\right)^2 H\left(\Phi_\theta(x,y)\right) + f_{\theta xy}^2\left(1 - H\left(\Phi_\theta(x,y)\right)\right)\right]}_{single\ segment\ term}\right.$$

$$\left.+\underbrace{\frac{k}{2}\left[\left(H\left(\Phi_\theta(x,y)\right) - H\left(\Phi_{\theta-1}(x,y)\right)\right)^2 + \left(H\left(\Phi_\theta(x,y)\right) - H\left(\Phi_{\theta+1}(x,y)\right)\right)^2\right]}_{term\ set\ 3}\right)$$

$$subject\ to \qquad f \geq 0$$

Equation 4-1

where f_θ, c_θ, and Φ_θ are the optimization variables. f_θ is the vectorized fluence map, c_θ is a value that f approaches within an aperture, and Φ_θ is the level set function, defined as positive where the aperture exists and negative elsewhere. The level set $\{(x,y)|\Phi_\theta(x,y) = 0\}$ describes the aperture boundary. Beam angles are indexed by θ, which ranges from 1 to n, and x and y are indices for a beamlet at a given θ. The fluence to dose transformation matrix is denoted by A, and the desired dose, d_0, is set as the prescription dose at the PTV and zero elsewhere. The diagonal weighting matrix, W, weights the structures of interest. The derivative matrices, D_1 and D_2, take the derivative of the fluence in both directions parallel and orthogonal to the MLC leaf movement. H is the Heaviside function

$$H(v) = \begin{cases} 1 & if\ v \geq 0 \\ 0 & if\ v < 0 \end{cases}$$

Essentially, $H\big(\Phi_\theta(x,y)\big)$ equals one inside the aperture and zero elsewhere. $f_{\theta xy}$ is a scalar value representing a single beamlet at a given beam angle θ and an x and y location on the beam, while f_θ is a vector of all the fluences at a specific beam angle. c_θ is a scalar quantity and only has one value per beam at a given time.

Intuitively, the first term is the dose fidelity term that attempts to push the final dose as close as possible to the desired dose. Term set 1, in equation 1, is the anisotropic total variation (TV) regularization, which has been shown to successfully encourage piecewise continuity on the fluence map. The TV regularization term considers the entire fluence map of the beam, so the term ultimately controls the segment size and shape, abating irregularities and holes in the aperture shape. Soft regulation of the minimal leaf gap and the max leaf interdigitation can be accomplished by independently adjusting the weightings λ_1 and λ_2, respectively. Term set 2 is pushing f towards c where the aperture is defined and zero elsewhere. Term set 3 encourages adjacent beam angles to be similar to regulate leaf movement between beam angles. For the 1st and nth θ, the $\Phi_{\theta-1}(x,y)$ and $\Phi_{\theta+1}(x,y)$ are equal to their respective $\Phi_\theta(x,y)$.

4.1.3 Results

Figure 4-1 shows the DVH of an LNG patient, comparing the comVMAT plans against the clnVMAT plans. It can be observed that comVMAT is able to better spare the OARs while maintaining a competitive PTV dosimetry. Figure 4-2 shows the dose wash for all of the patients. The comVMAT gave a much heavier weighting to some selective beams, while the clnVMAT overall spreads the distribution of fluence intensities more evenly among the beams, giving rise to a less angular modulated dose distribution pattern and greater dose to OARs. For example, for the GBM patient, the comVMAT plan was able to entirely avoid the brainstem, while the clnVMAT plan covers the brainstem with at least 2.5 Gy of dose.

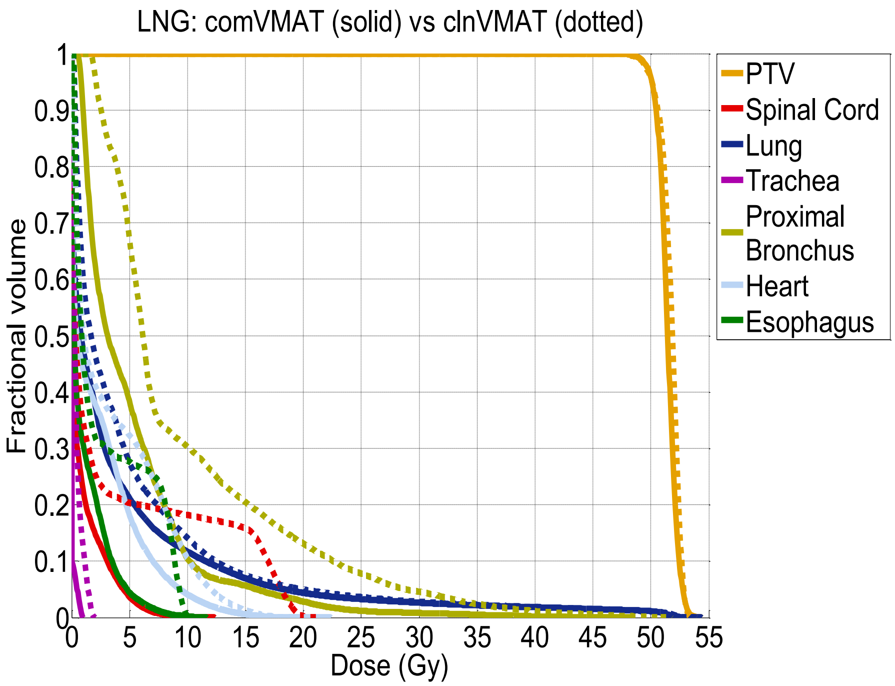


Figure 4-1 DVHs of an LNG patient comparing the comVMAT and clnVMAT plans.

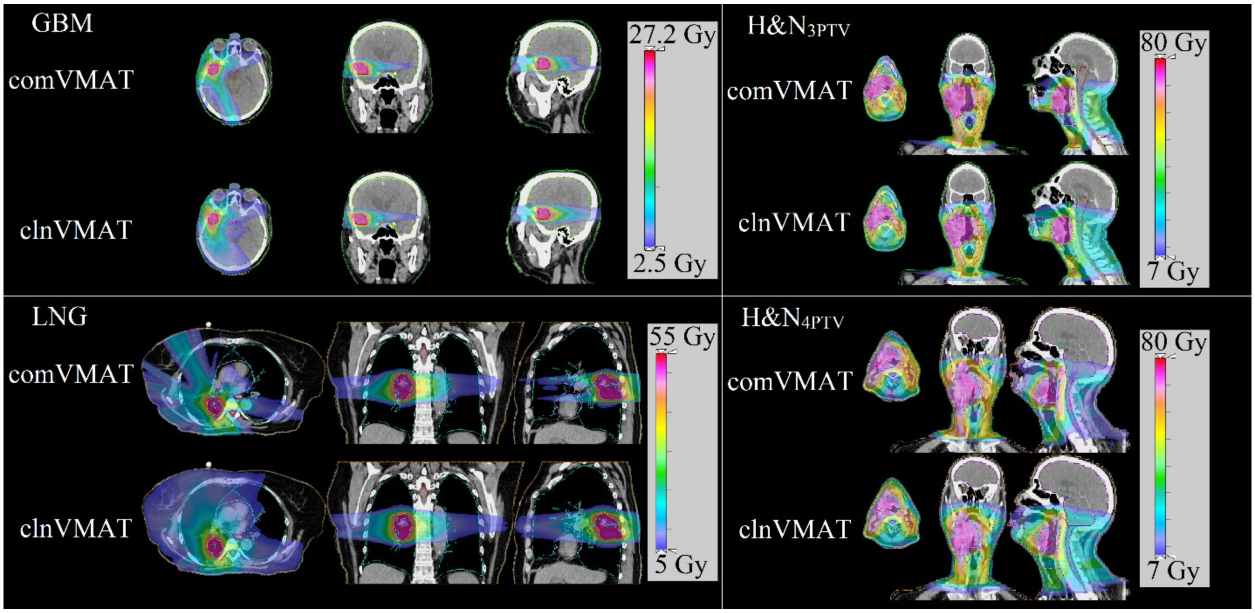


Figure 4-2 Dose washes for all patients. The low dose cutoff for viewing is set at 10% of the prescription dose.

4.1.4 Summary

We proposed an integrated VMAT framework that improves dosimetry compared with clinical VMAT. This study paves the foundation for the following sections in this chapter.

4.2 Single-Arc VMAT optimization for Dual-Layer MLC

4.2.1 Introduction

To overcome the PSO limitations, in the last section, we developed a full-angular-resolution direct aperture optimization (DAO) method for VMAT, which solves the VMAT problem for an entire arc without any interpolations. The previously proposed DAO VMAT using a single arc was shown to outperform the standard commercial algorithm (PSO method) using two arcs. With the same target coverage, the single-arc DAO VMAT plan enhanced OAR sparing, improved dose conformality, and reduced delivery time compared with the two-arc PSO VMAT plan.

The published VMAT algorithms thus far are developed for single-layer multi-leaf collimator (SLMLC). As an alternative to SLMLC, the dual-layer multi-leaf collimator (DLMLC) [95–97] with “stacked and staggered” leaves has recently attracted renewed interest due to two distinct advantages. First, the leakage dose is substantially reduced. Second, the MLC leaves can be wider and simpler without needing the tongue-and-groove structure, making them easier to fabricate. Two new commercial medical Linacs, Halcyon (Varian Medical Systems) and MRIdian (ViewRay) have adopted DLMLC. Although the general interest in these systems and the time dual-layer MLC has been made available, there has not been a VMAT method specifically developed for it. The dosimetric potential of DLMLC is hampered by the coupling effect of two MLC layers, which is not effectively managed in the existing PSO interpolation approach. As a result, more arcs have been shown necessary to achieve comparable dosimetry to a VMAT plan by single layer higher resolution MLC [98]. In this study, we proposed a non-PSO single-arc VMAT optimization framework that simultaneously solves for both MLC layers and the entire arc in full resolution, using an alternating optimization approach that have been investigated in our previous VMAT optimization studies [35,38,39].

4.2.2 Methods

Two available DLMLC platforms differ by physical dimensions, mechanical parameters, and whether they are double-focused. However, from the optimization perspective, without losing generality, we use the Halcyon properties, which currently support VMAT delivery. In contrast to the standard MLC using single layer 5 mm MLC (Figure 4-3B, SLMLC-5mm), the DLMLC utilizes two layers of MLC banks on both sides, with 10mm leaf width projected at the isocenter (Figure 4-3A, DLMLC-10mm). The two MLC layers are stacked and staggered by half of the leaf width to provide more sophisticated modulation than SLMLC with the same leaf width (Figure 4-3C, SLMLC-10mm). The extension of each leaf is usually restricted to a step size matching the leaf width to simplify the computational burden during planning. As an additional and readily achievable configuration, the SLMLC with 10 mm leaf width could employ extensions with a smaller step size of 5 mm (Figure 4-3D, SLMLC-10mm-5mm) to form rectangular beamlets. Figure 4-3 shows the MLC aperture that best spares an ellipse-shaped OAR while keeping high PTV coverage of a C-shaped target. The DLMLC-10mm design forms the same aperture as the SLMLC-5mm design, while the SLMLC-10mm and

SLMLC-10mm-5mm inevitably expose more OAR to maintain the PTV coverage demonstrating DLMLC's capability to improve modulation resolution.

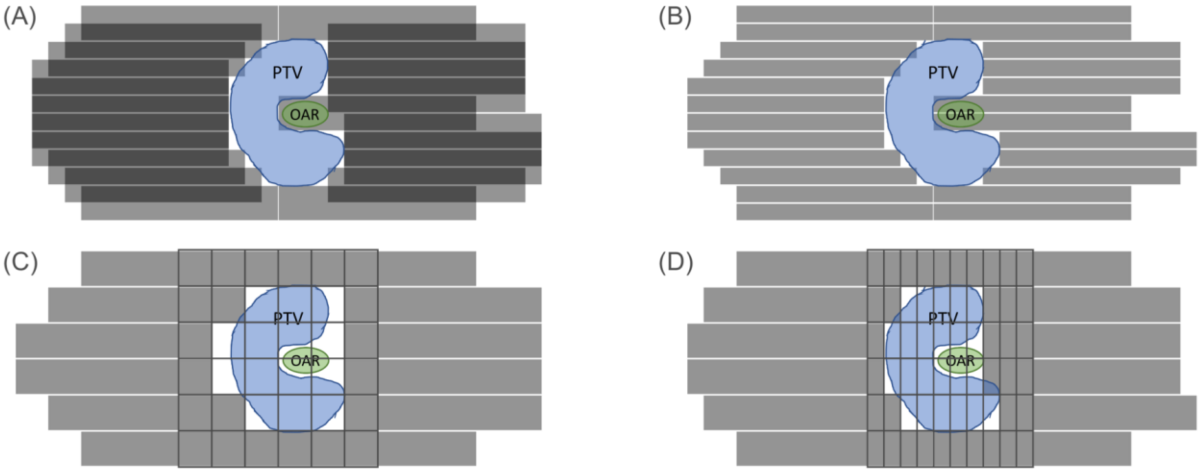


Figure 4-3 Demonstration of (A) DLMLC with 10mm leaf width (DLMLC-10mm), (B) SLMLC with 5mm leaf width (SLMLC-5mm), (C) SLMLC with 10mm leaf width (SLMLC-10mm), (D) SLMLC with 10mm leaf width and 5mm leaf step size (SLMLC-10mm-5mm). The grids on (C) and (D) represent the achievable beamlets.

4.2.2.1 Deliverability of Dual-Layer MLC

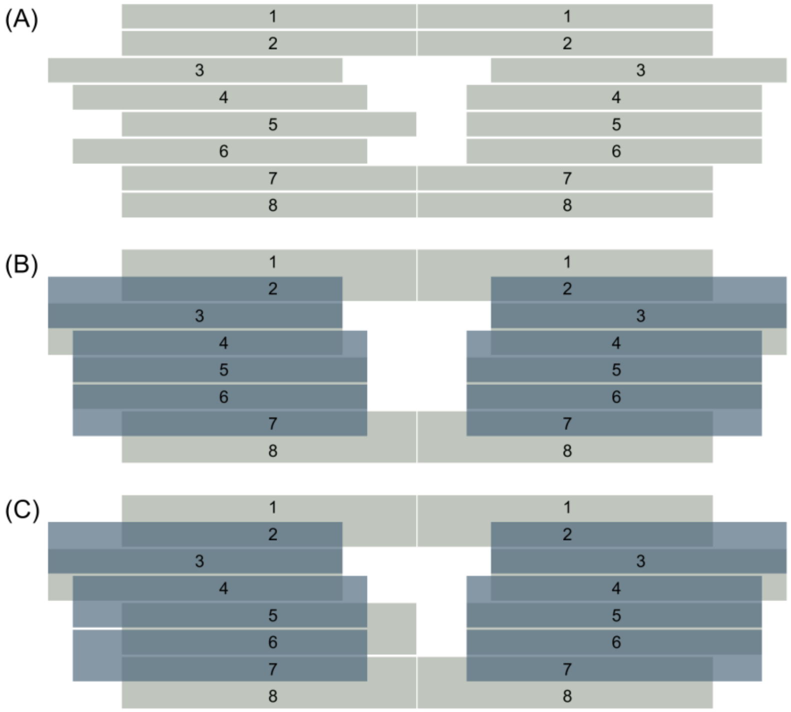


Figure 4-4 Comparison of the SLMLC with 5mm leaf width (SLMLC-5mm) and DLMLC with 10mm leaf width (DLMLC-10mm). The 5th MLC leaf position is extended farther than both neighboring leaves, resulting in an aperture shape that is undeliverable using DLMLC-10mm. (B) and (C) are examples of DLMLC deliverable aperture shapes that are closest to the shape of aperture (A).

To understand the coupling effect of the dual MLC layers, we first examine the deliverability of DLMLC. Deliverable apertures on SLMLC-10mm or SLMLC-10mm-5mm are deliverable on DLMLC-10mm since one of the dual MLC layers could be kept open while moving the other layer. Although any DLMLC-10mm deliverable patterns are also deliverable using SLMLC-5mm, the converse is not true. A necessary and sufficient condition for an SLMLC-5mm aperture to be also deliverable using DLMLC-10mm is that there is no SLMLC-5mm leaf farther than its adjacent MLC leaves towards the opposing bank. Figure 4-4 shows an example of SLMLC-5mm deliverable aperture (Figure 4-4A) that is not achievable using DLMLC-10mm, since the 5th MLC leaf on the left side protrudes farther than both of its neighboring leaves. Two of the most similar apertures achievable on DLMLC-10mm are shown in Figure 2B and Figure 2C, where either the 5th MLC leaf is moved back to match with the 4th leaf or the 6th leaf (Figure 4-4B), or the 6th leaf is moved forward to match with the 5th leaf (Figure 4-4C). However, either configuration would result in a different fluence map and subsequently different delivered dose.

4.2.2.2 Direct Aperture Optimization Formulation

Consider the constraints, the DLMLC-10mm VMAT optimization problem is formulated as

$$\underset{\{f_\theta, c_\theta, u_{1,\theta}, u_{2,\theta}\}_{\theta=0}^{n}}{\text{minimize}} \quad \underbrace{\frac{1}{2}\left\| W\left(\left(\sum_\theta A_\theta f_\theta\right) - d_0\right)\right\|_2^2}_{\text{fidelity term}} + \sum_\theta \underbrace{\left(\lambda_x \|D_x f_\theta\|_1 + \lambda_y \|D_y f_\theta\|_1\right)}_{\text{anisotropic TV term on } f}$$

$$+ \sum_\theta \underbrace{\frac{\gamma}{2}\left(\|\operatorname{diag}(u_\theta)^{1/2}(f_\theta - c_\theta)\|_2^2 + \|\operatorname{diag}(1-u_\theta)^{1/2} f_\theta\|_2^2\right)}_{\text{single segment term}}$$

$$\text{subject to} \qquad f_\theta \geq 0, c_\theta \geq 0, u_{i,\theta} \in \{0,1\}^{n_x n_y}, \qquad i = 1,2 \quad \theta = 1,2,\dots,n_\theta$$

$$u_\theta = u_{1,\theta} \text{ AND } u_{2,\theta}$$

$$\left(u_{i,\theta}, u_{i,\theta+1}\right) \in S, \qquad i = 1,2 \quad \theta = 1,2,\dots,n_\theta$$

Equation 4-2

where the notations for the variables and data are summarized in Table 4-1.

$u_{1,\theta}$ and $u_{2,\theta}$ describe the aperture of upper and lower MLC layers with 10 mm leaf. $u_{1\theta}$ (or $u_{2\theta}$) is 1 where the aperture exists, and 0 elsewhere. The effective aperture u_θ has 5mm resolution combining both MLC layers. The effective aperture is open if and only if the corresponding beamlet is open on both MLC layers and is closed if the corresponding beamlet is closed on at least one of the MLC layers, i.e., u_θ is 1 where both $u_{1\theta}$ and $u_{2\theta}$ equal to 1, and 0 elsewhere. Note that although the resultant effective aperture u_θ has 5mm resolution, it is readily deliverable on DLMLC-10mm, as it is defined by two coarse apertures $u_{1\theta}$ and $u_{2\theta}$. The same effective aperture u_θ can be mapped to more than one configuration of $u_{1\theta}$ and $u_{2\theta}$, affording more options to meet the MLC leaf speed constraints. The set S is then defined to include all achievable apertures that satisfy the MLC speed constraint. If two adjacent beams $u_{i,\theta}$ and $u_{i+1,\theta}$ have substantially different apertures such that it is impossible for the MLC leaves to travel from $u_{i,\theta}$ to $u_{i+1,\theta}$ within the time limit of each control point, the aperture configuration $\left(u_{i,\theta}, u_{i,\theta+1}\right)$ will be excluded in set S. Note that the speed constraint applies to the two layers separately.

The fidelity term attempts to find the optimal fluence map f_θ by penalizing the difference between the planning dose and the ideal dose d_0, in a least-square form, where the diagonal weighting matrix W regulates the priorities of the structure of interest. The anisotropic total variation (TV) term encourages the piecewise continuity of the fluence map f_θ. The amount of smoothness is controlled by λ_x and λ_y, in the direction parallel and orthogonal to the MLC leaf motion direction, respectively.

The single segment term simplifies the final fluence map f_θ to contain only one segment at each control point, which corresponds to the effective aperture u_θ. It pushes the fluence map f_θ towards the uniform intensity value c_θ where the effective aperture u_θ is 1, and towards 0 elsewhere, encouraging the fluence map to reach a constant intensity level within the effective aperture and 0 outside the effective aperture. By gradually increasing the weighting parameter γ in the optimization, the algorithm attains the single segment constraint for VMAT.

Notation	Type	Description
		Indices
$\boldsymbol{\theta}$	Index	Index for gantry angle, $\theta = 1,2,\ldots,n_\theta$
$\boldsymbol{i}$	Index	Index for MLC layer, $i = 1,2$
$\boldsymbol{x}$	Index	Index for beamlet in the direction parallel to the MLC leaf movement, $x = 1,2,\ldots,2n_x$
$\boldsymbol{y}$	Index	Index for beamlet in the direction orthogonal to the MLC leaf movement, $y = 1,2,\ldots,2n_y$
		Optimization Variables
$\boldsymbol{f_\theta}$	Vector	High-resolution vectorized fluence map $f_\theta \in \mathbb{R}^{4n_x n_y}$ for the candidate beam θ
$\boldsymbol{c_\theta}$	Vector	Intensity value $c_\theta \in \mathbb{R}^{4n_x n_y}$ that f_θ approaches within the aperture of beam θ
$\boldsymbol{u_{i,\theta}}$	Vector	Aperture variable $u_{i,\theta} \in \{0,1\}^{n_x n_y}$ for MLC layer i and beam θ Defined as 1 where the aperture exists and 0 elsewhere
$\boldsymbol{u_\theta}$	Vector	Effective aperture variable $u_\theta \in \{0,1\}^{4n_x n_y}$ for beam θ Defined as 1 where the aperture exists on both MLC layers and 0 elsewhere (logical "AND" of $u_{1,\theta}$ and $u_{2,\theta}$)
		Other data
$\boldsymbol{A_\theta}$	Matrix	Fluence to dose transformation matrix for beam at gantry angle θ

$\boldsymbol{W}$	Matrix	Diagonal weighting matrix, with weightings for structures of interest as diagonal elements
$\boldsymbol{d_0}$	Vector	Ideal dose with the prescription dose at the PTV and zero elsewhere
$\boldsymbol{D_x}$	Matrix	Derivative matrix in the direction parallel to the MLC leaf movement
$\boldsymbol{D_y}$	Matrix	Derivative matrix in the direction orthogonal to the MLC leaf movement
AND	Operator	Logic operator that outputs 1 if and only if both inputs equal to 1, otherwise 0
$\boldsymbol{S}$	Set	A set of all achievable apertures that satisfy MLC mechanical constraint

Table 4-1 Notations and data structures used in this study.

4.2.2.3 Algorithm

Similar to our previous studies [35,38,39], the optimization problem in equation (1) is solved via alternating optimization by fixing two of the three optimization variables and optimizing with respect to the other variable. In each iteration, the optimization algorithm runs through 3 modules and optimizes with respect to the fluence map f_θ, the aperture $u_{i\theta}$, and the intensity value c_θ, respectively. The process is repeated until f_θ converges to $c_\theta \odot u_\theta$, where $\odot$ is the Hadamard product, implying that constant intensity is achieved within each optimized aperture.

4.2.2.3.1 Module 1: Optimize the fluence map f_θ

In module 1, the aperture $u_{i\theta}$ and the intensity value c_θ are kept constant, and the equivalent optimization problem with respect to the fluence map f_θ is a convex optimization problem:

$$\begin{aligned} \underset{\{f_\theta\}_{\theta=0}^{n}}{\text{minimize}} \quad & \frac{1}{2}\left\| W\left(\left(\sum_\theta A_\theta f_\theta\right) - d_0\right)\right\|_2^2 + \sum_\theta \left(\lambda_x \|D_{x\theta} f_\theta\|_1 + \lambda_y \left\|D_{y\theta} f_\theta\right\|_1\right) \\ & + \sum_\theta \frac{\gamma}{2} \left\|\operatorname{diag}(u_\theta)^{1/2}(f_\theta - c_\theta)\right\|_2^2 + \left\|\operatorname{diag}(1-u_\theta)^{1/2} f_\theta\right\|_2^2 \\ \text{subject to} \quad & f_\theta \geq 0, \quad \theta = 1,2,\ldots,n_\theta \end{aligned}$$

Equation 4-3

The optimization problem in Equation 4-3 can be efficiently solved by FISTA.

4.2.2.3.2 ***Module 2: Optimize the aperture $\mathbf{u_\theta}$***

In module 2, the fluence map f_θ and the intensity value c_θ are kept constant. The intensity value is estimated as the average intensity value inside the aperture from the last iteration. The algorithm attempts to find the optimal effective aperture u_θ, jointly defined by the two MLC layers that best fit the current fluence map f_θ and an estimated intensity value c_θ, while ensuring the deliverability constraints. The equivalent optimization problem with respect to the aperture u_θ is:

$$\underset{\{u_\theta, u_{1,\theta}, u_{2,\theta}\}_{\theta=0}^{n}}{\text{minimize}} \qquad \sum_\theta \langle q_\theta, u_\theta \rangle,$$

$$\text{subject to} \qquad u_{i,\theta} \in \{0,1\}^{n_x n_y},\ \ \left(u_{i,\theta}, u_{i,\theta+1}\right) \in S,\ \ i = 1,2 \quad \theta = 1,2, \dots, n_\theta$$

$$u_\theta = u_{1,\theta} \text{ AND } u_{2,\theta}.$$

where q_θ is a vector with elements $q_{\theta xy}$ defined as:

$$q_{\theta xy} = \left(f_{\theta xy} - c_\theta\right)^2 - f_{\theta xy}^2 = c_\theta\left(c_\theta - 2f_{\theta xy}\right).$$

Equation 4-4

$q_{\theta xy}$ can be interpreted as the cost associated with a beamlet. The beamlet has a positive cost if $f_{\theta xy} < \frac{c_\theta}{2}$ and a negative cost if $f_{\theta xy} > \frac{c_\theta}{2}$. Dosimetrically favorable beamlets, which have a higher intensity value in the optimization, present negative beamlet costs and therefore are more likely to be open.

Note that the objective function is linear in u_θ, where u_θ is either 0 or 1. The minimization of the objective function is equivalent to a simplified travelling salesman problem on a directed graph $\mathrm{G} = [N, E]$ with a rectangular structure. The graph is designed to contain n_u by n_θ nodes, where n_u is the number of all possible configurations of the aperture u_θ, and

n_θ is the number of control points. The graph is associated with a node cost $\mathrm{NC}(u_\theta, \theta)$ for an aperture u_θ at gantry angle θ, and the edge cost $\mathrm{EC}\big((u_{\theta 1}, \theta_1), (u_{\theta 2}, \theta_2)\big)$ that connects $(u_{\theta 1}, \theta_1)$ and $(u_{\theta 2}, \theta_2)$. The node cost is defined as

$$\mathrm{NC}\big((u_\theta, \theta)\big) = \begin{cases} \langle q_\theta, u_\theta \rangle, & \exists u_{1\theta}, u_{2\theta},\ s.t.\, u_\theta = u_{1\theta} \text{ AND } u_{2\theta} \\ +\infty, & \text{otherwise} \end{cases},$$

which is finite if and only if u_θ is DLMLC deliverable. The edge cost is defined as

$$\mathrm{EC}\big((u_{\theta 1}, \theta_1), (u_{\theta 2}, \theta_2)\big) = \begin{cases} 0, & (u_{\theta 1}, u_{\theta 2}) \in S, \theta_2 - \theta_1 = \Delta\theta \\ +\infty, & \text{otherwise} \end{cases},$$

Equation 4-5

where $\Delta\theta$ is the gantry angle between two adjacent control points. The edge cost is finite if and only if the edge connects two adjacent beams, whose apertures satisfying MLC leaf travelling speed constraints, and is directed towards the next control point.

The traveling salesman problem finds the shortest path on the graph, leading to a solution with minimum total node and edge costs. Therefore, apertures that are not DLMLC deliverable or violate the MLC leaf speed constraints will be automatically excluded. The exact solution to the graph optimization problem requires the calculation of all possible aperture configurations. To reduce the computation complexity, we adopted a heuristic approach. Firstly, the best aperture $u_{\theta,best}$ without deliverability constraints is computed by switching on all beamlets that are associated with a negative $q_{\theta xy}$, and switching off all other beamlets. Undeliverable apertures are identified on $u_{\theta,best}$ and modified by approximating it with an SLMLC-5mm deliverable aperture with the lowest total opened beamlet cost. The modification is achieved by either filling in the holes inside the aperture or switching on some closed beamlets. The modified aperture, denoted by $u_{\theta,SL}$, is deliverable using SLMLC-

5mm. SLMLC-5mm deliverable apertures that are not deliverable using DLMLC-10mm are identified on $u_{\theta,SL}$ by looking for any leaf that is positioned farther than its adjacent leaves. The best fit $u_{\theta,DL}$ is computed by moving the violating leaves or its neighboring leaves to ensure DLMLC-10mm deliverability constraint while achieving the lowest total opened beamlet cost. Note that although a few approximations are made to the original aperture $u_{\theta,best}$ in doing so, the modified apertures still correspond to the lowest total beamlet costs under the set of deliverability constraints, i.e., the SLMLC-deliverability constraint for $u_{\theta,SL}$ and the DLMLC-deliverability constraint for $u_{\theta,DL}$. To minimize leakage through one layer of MLC, the leaf positions of both MLC layers are extended as far as possible without blocking any opened beamlets defined by $u_{\theta,DL}$. The aperture defined by the coarse MLC layers are denoted by $u_{i\theta,DL}$, for the ith MLC layer.

Any aperture configuration $\left((u_{i\theta,DL},\theta),(u_{i(\theta+\Delta\theta),DL},\theta+\Delta\theta)\right)$ that violates the MLC speed limit constraint is identified based on $u_{i\theta,DL}$. Exhaustive search is performed on the identified apertures to minimize the total opened beamlet cost while ensuring the MLC traveling constraints with the control point and its adjacent beams $(\theta-\Delta\theta,\theta,\theta+\Delta\theta)$. The process is repeated until all beams satisfy the MLC leaf speed constraints. Although the exhaustive search could be computationally expensive if performed on all leaves, in practice, violation of the leaf speed constraints infrequently happens on only a few leaves and control points. Therefore, it only takes seconds for the whole optimization process with respect to aperture variable $u_{i\theta}$. The resultant apertures defined by the two MLC layers satisfy both the MLC leaf speed constraint and the DLMLC-10mm deliverability constraint.

4.2.2.3.3 ***Module 3: Optimize the intensity value c_θ***

In module 3, the aperture u_θ and the fluence map f_θ are kept constant. With known aperture u_θ, it is straightforward to find the intensity value c_θ that achieves a final dose that is closest to the ideal dose. The optimization subproblem with respect to the intensity value c_θ is

$$\underset{\{c_\theta\}_{\theta=0}^{n}}{\text{minimize}} \quad \frac{1}{2}\left\| W\left(\left(\sum_\theta A_\theta(u_\theta \odot c_\theta)\right) - d_0\right)\right\|_2^2$$

$$\text{subject to} \quad c_\theta \geq 0, \quad \theta = 1,2,\ldots,n_\theta,$$

Equation 4-6

which could be solved by FISTA efficiently.

4.2.2.4 Direct Aperture Optimization Formulation for Simultaneous Integrated Boost

An interesting and relevant challenge that tests the MLC resolution is to perform simultaneous integrated boost on a small boost volume. For this purpose, the DLMLC-10mm VMAT optimization for Simultaneous Integrated Boost (SIB) is formulated as

$$\begin{aligned}
\underset{\{f_\theta, c_\theta, u_{1,\theta}, u_{2,\theta}\}_{\theta=0}^{n}}{\textbf{minimize}} \quad & \underbrace{\frac{1}{2}\left\| W\left(\left(\sum_\theta A_\theta f_\theta\right) - d_0\right)\right\|_2^2}_{\text{fidelity term}} - \underbrace{w\mathbf{1}^T \sum_\theta A_{\theta,SIB} f_\theta}_{\text{boost term}} \\
& + \lambda_{\text{SIB}} \underbrace{\left(\max\left(\sum_\theta A_{\theta,\text{SIB}} f_\theta\right) - \min\left(\sum_\theta A_{\theta,SIB} f_\theta\right)\right)}_{\text{homogeneity term}} \\
& + \sum_\theta \underbrace{\left(\lambda_x \|D_{x\theta} f_\theta\|_1 + \lambda_y \left\|D_{y\theta} f_\theta\right\|_1\right)}_{\text{anisotropic TV term on } f} \\
& + \sum_\theta \underbrace{\frac{\gamma}{2}\left\|\text{diag}(u_\theta)^{1/2}(f_\theta - c_\theta)\right\|_2^2 + \left\|\text{diag}(1-u_\theta)^{1/2} f_\theta\right\|_2^2}_{\text{single segment term}}, \\
\textbf{subject to} \quad & f_\theta \geq 0, c_\theta \geq 0, u_{i,\theta} \in \{0,1\}^{n_x n_y}, \quad i = 1,2 \quad \theta = 1,2,\dots,n_\theta \\
& u_\theta = u_{1,\theta} \text{ AND } u_{2,\theta} \\
& \left(u_{i,\theta}, u_{i,\theta+1}\right) \in S, \quad i = 1,2, \quad \theta = 1,2,\dots,n_\theta
\end{aligned}$$

Equation 4-7

where there are two additional terms: boost term and homogeneity term, compared with Equation 4-2. The boost term attempts to escalate the dose as great as possible within the boost volume by maximizing its mean dose. $A_{\theta,SIB}$ is the fluence to dose transformation matrix within the boost area for the beam at gantry angle θ. The parameter w regulates the weighting of the boost structure. The homogeneity term controlled by λ_{SIB} encourages the dose uniformity within the boost area.

4.2.2.5 Evaluation

The proposed optimization algorithm was tested on four patients, including a glioblastoma multiforme case (GBM), a lung cancer case (LNG), a prostate cancer case (PRT), and a rectum cancer case with SIB volume (REC-SIB), where the boost volume is determined by diffusion MRI and validated by pathology [99]. The prescription doses and PTV volumes for all patients and the volume of the boost area for the SIB case can be found in Table 4-2.

	Prescription Dose (Gy)	PTV Volume (cc)	Boost Volume (cc)
GBM	25	6.36	
LNG	50	9.33	
PRT	40	127.02	
REC-SIB	40	1145.52	0.14

Table 4-2 Prescription doses and PTV volumes for all patients, and the volume of the boost area for the SIB case.

The dose matrix was obtained using a convolution/superposition dose calculation algorithm with 6 MV x-ray polyenergetic kernels as described in our previous publications [100], calculated for 180 beams with 2-degree gantry angle spacing. The tissue densities are provided by CT images for the first three patients and assumed 1 g/cm^3 for the MR-based REC-SIB treatment. The couch angle is set at 0 degrees, and the collimator at 45 degrees, following the International Electrotechnical Commission (IEC) convention. The beamlet resolution for dose calculation was 5×5 mm^2 for SLMLC-5mm and DLMLC-10mm, 10×10 mm^2 for SLMLC-10mm, and 5×10 mm^2 for SLMLC-10mm-5mm. The dose array resolution was 2.5×2.5×2.5 mm^3.

Evaluation with respect to PTV includes PTV D95, D98, D99, D2 (defined as the dose received by at least 95%, 98%, 99%, and 2% of the target volume, respectively), PTV homogeneity (defined as $\frac{\mathrm{D95}}{\mathrm{D5}}$), and dose conformity (defined as the ratio between the PTV volume receiving 100% or more of the prescription dose and the PTV volume). For OARs, the maximum dose (Dmax) and mean dose (Dmean) were evaluated, where Dmax is defined as D2, the dose at 2% of the structure volume, recommended by the ICRU-83 report [101]. The integral dose, defined as the volume integral of the dose deposited in the patient, and the R50, defined as the 50% isodose volume divided by the target volume, were also obtained to quantify overall dose spillage and high dose spillage, respectively.

4.2.3 Results

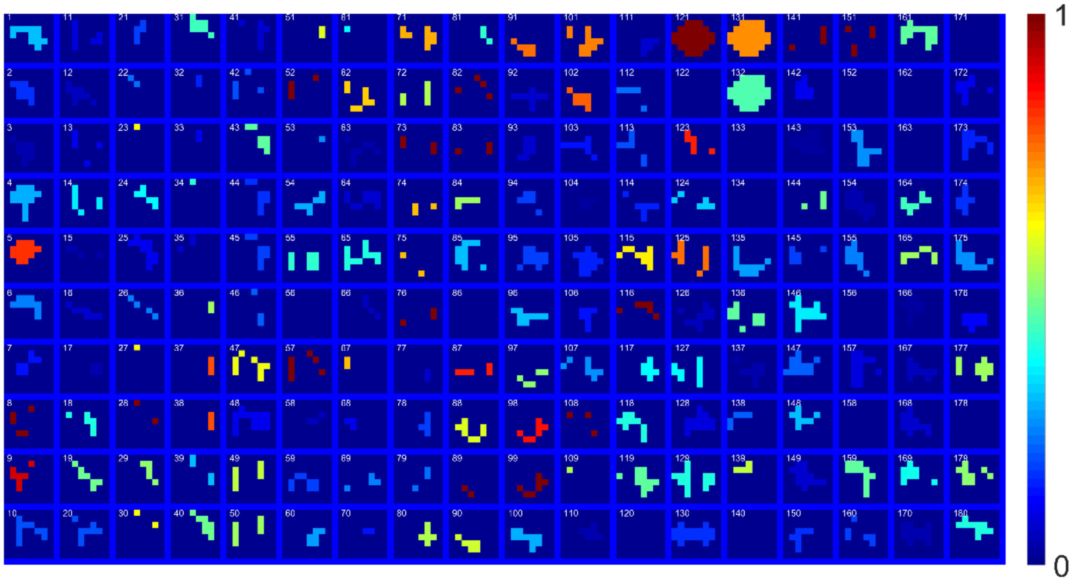


Figure 4-5 Fluence map of the DLMLC-10mm plan for the lung cancer patient. The beamlet resolution is 5mm, while the fluence map is still deliverable with DLMLC-10mm. MLC leaf direction is vertical.

Figure 4-5 shows the fluence map of the DLMLC-10mm plan for the lung cancer patient, including 180 control points. The proposed DLMLC VMAT optimization algorithm optimizes all beams simultaneously and produces 5mm-resolution apertures that are deliverable using DLMLC-10 mm. In this study, the maximal MLC leaf travel is set at 10 mm between two adjacent beams. For the MLC leaves traveling at the maximal speed of 50 mm/s (Halcyon white paper), 0.2 seconds is needed between control points. The estimated delivery time is then 36 seconds for the GBM and LUNG cases and 90 seconds for the PRT case.

Figure 4-6 shows the DVHs of the DLMLC-10mm plan and the SLMLC plans for all patients. PTV D95 is normalized to the prescription dose. The PTV statistics are matched in all plans for fair OAR dose comparison. The DLMLC-10mm and the SLMLC-5mm plans are nearly indistinguishable. With the same leaf width, the DLMLC-10mm plan OAR sparing is superior to both the SLMLC-10mm-5mm and the SLMLC-10mm plans for all patients. In the GBM case, the dose to the brainstem, chiasm, and right optic nerve was substantially reduced. In the LNG case, hot spots are significantly reduced in dose-limiting organs such as the major vessel, proximal bronchus, and heart. In the PRT case, the dose to the rectum, penile bulb, and seminal vesicle was effectively controlled. For the boost case, with comparable or better OAR sparing, the DLMLC-10mm plan achieved greater boost dose than both the SLMLC-10mm-5mm and SLMLC-10mm plans. The detailed dose statistics of the boost case is summarized in Table 4-3.

SIB statistics	**D98 (Gy)**	**D2 (Gy)**
DLMLC-10mm	62.21	74.55
SLMLC-5mm	62	79.04
SLMLC-10mm	52.22	59.44
SLMLC-10mm-5mm	60	73.31

Table 4-3 Dose statistics of the boost area for the SIB case.

Figure 4-7 shows the dose distribution of the DLMLC-10mm and the SLMLC plans for all patients. Compared with SLMLC with the same leaf width, DLMLC-10mm reduced the overall OAR dose, especially those adjacent to PTV, while maintaining the same PTV coverage. Table 4-4 shows the quantitative PTV statistics, R50, and Integral Dose. The D95 were normalized to the prescription dose for all plans. The PTV homogeneity, PTV D2, and PTV D98 are comparable across all plans. The DLMLC-10mm reduced R50 by 30.7% and 10.0% compared with the SLMLC-10mm plan and the SLMLC-10mm-5mm plan respectively, indicating a remarkable improvement in dose compactness.

Table 4-6 shows the OAR statistics for all patients. With the same target coverage, compared with the SLMLC-5mm plan, the DLMLC-10mm plan slightly increased the OAR max and mean doses by 0.35% and 0.37% of the prescription dose. Even the largest difference of single OAR max and mean dose is within 2.1% and 0.9% of the prescription dose, showing comparable

OAR sparing across the DLMLC-10mm plan and the SLMLC-5mm plan. Compared with the SLMLC-10mm plan, the DLMLC-10mm plan reduced the OAR max and mean doses by 5.79% and 4.18% of the prescription dose on average, and the single largest sparing in OAR max and mean dose is up to 27.6% and 14.4% of the prescription dose. Compared with the SLMLC-10mm-5mm plan, the DLMLC-10mm plan reduced the OAR max and mean doses by 3.7% and 2.1% of the prescription dose on average, and the single largest sparing in OAR max and mean dose is up to 20.9% and 10.2% of the prescription dose.

Table 4-5 shows the optimization time for all plans. The DAO takes from 10 minutes to 2 hours for typical VMAT plans, depending on the PTV size. The computation bottleneck is the multiplication with the system matrix. Down sampling approaches, which were not implemented in our experiments, are expected to substantially accelerate the optimization.

Patient Case			DLMLC-10mm	SLMLC-5mm	SLMLC-10mm	SLMLC-10mm-5mm
PTV Statistics	D2 (Gy) (SLMLC plans) – (DLMLC-10mm plan)	GBM		-0.01	-0.24	-0.15
		LNG		0.09	-0.47	0.02
		PRT		-0.07	-0.25	-0.2
		REC-SIB		-0.29	0.21	-0.04
	D98 (Gy) (SLMLC plans) – (DLMLC-10mm plan)	GBM		-0.03	-0.04	0.01
		LNG		-0.05	0.27	-0.12
		PRT		-0.13	0.05	0.31
		REC-SIB		0.54	-0.14	0.32
	Homogeneity	GBM	0.94	0.94	0.94	0.94
		LNG	0.95	0.95	0.94	0.95
		PRT	0.94	0.94	0.94	0.93
		REC-SIB	0.91	0.9	0.91	0.91
R50		GBM	4.7	4.81	7.99	4.94
		LNG	2.78	2.85	5.51	3.57
		PRT	3.04	3.14	3.41	3.18
		REC-SIB	2.69	2.74	3.43	2.94
Integral Dose (Gy·m^3)		GBM	2.46	2.55	3.89	3.07
		LNG	12.58	12.54	18.59	14.62
		PRT	64.84	65.11	71.2	66.9
		REC-SIB	189.91	190.84	209.7	194.34

Table 4-4 PTV statistics, R50, and Integral Dose for all patients.

Optimization time (min)	DLMLC-10mm	SLMLC-5mm	SLMLC-10mm	SLMLC-10mm-5mm
GBM	8	8	4	6
LNG	32	32	18	24
PRT	114	96	64	91
REC-SIB	238	231	95	153

Table 4-5 Optimization time for all plans.

Patient Case			GBM	LNG	PRT	SEC-SIB
OAR dose sparing: (SLMLC plans) – (DLMLC-10mm plan) (Gy)	Dmax Largest Difference	SLMLC-5mm	0.3 Lens, Right	0.25 Heart	0.82 Femur, Right	0.51 Bowel
		SLMLC-10mm	4.86 Optic Nerve, Right	13.81 Major Vessels	8.26 Penile Bulb	2.32 Femur, Right
		SLMLC-10mm-5mm	5.24 Optic Nerve, Right	7.05 Major Vessels	2.63 Femur, Left	1.12 Femur, Left
	Dmax Average Difference	SLMLC-5mm	-0.22	-0.53	0.2	0.03
		SLMLC-10mm	2.33	4.14	1.6	0.61
		SLMLC-10mm-5mm	1.63	2.13	1.12	0.49
	Dmean Largest Difference	SLMLC-5mm	0.17 Lens, Right	0.14 Spinal Cord	0.34 Femur, Left	-0.07 Femur, Left
		SLMLC-10mm	3.6 Optic Nerve, Right	3.35 Major Vessels	4.14 Rectum	1.34 Bladder
		SLMLC-10mm-5mm	2.55 Optic Nerve, Right	1.71 Major Vessels	1.82 Rectum	1.09 Femur, Left
	Dmean Average Difference	SLMLC-5mm	0.03	-0.03	0.07	-0.7
		SLMLC-10mm	1.6	1.08	2.53	0.73
		SLMLC-10mm-5mm	1.11	0.63	1.03	0.06

Table 4-6 OAR mean and maximum dose sparing differences between SLMLC plans and the DLMLC-10mm plan for all patients. The rows labeled "Largest Values" represents the largest amount of dose sparing difference achieved among all OARs, and the corresponding OAR. "Average values" represents the average dose sparing difference among all OARs.

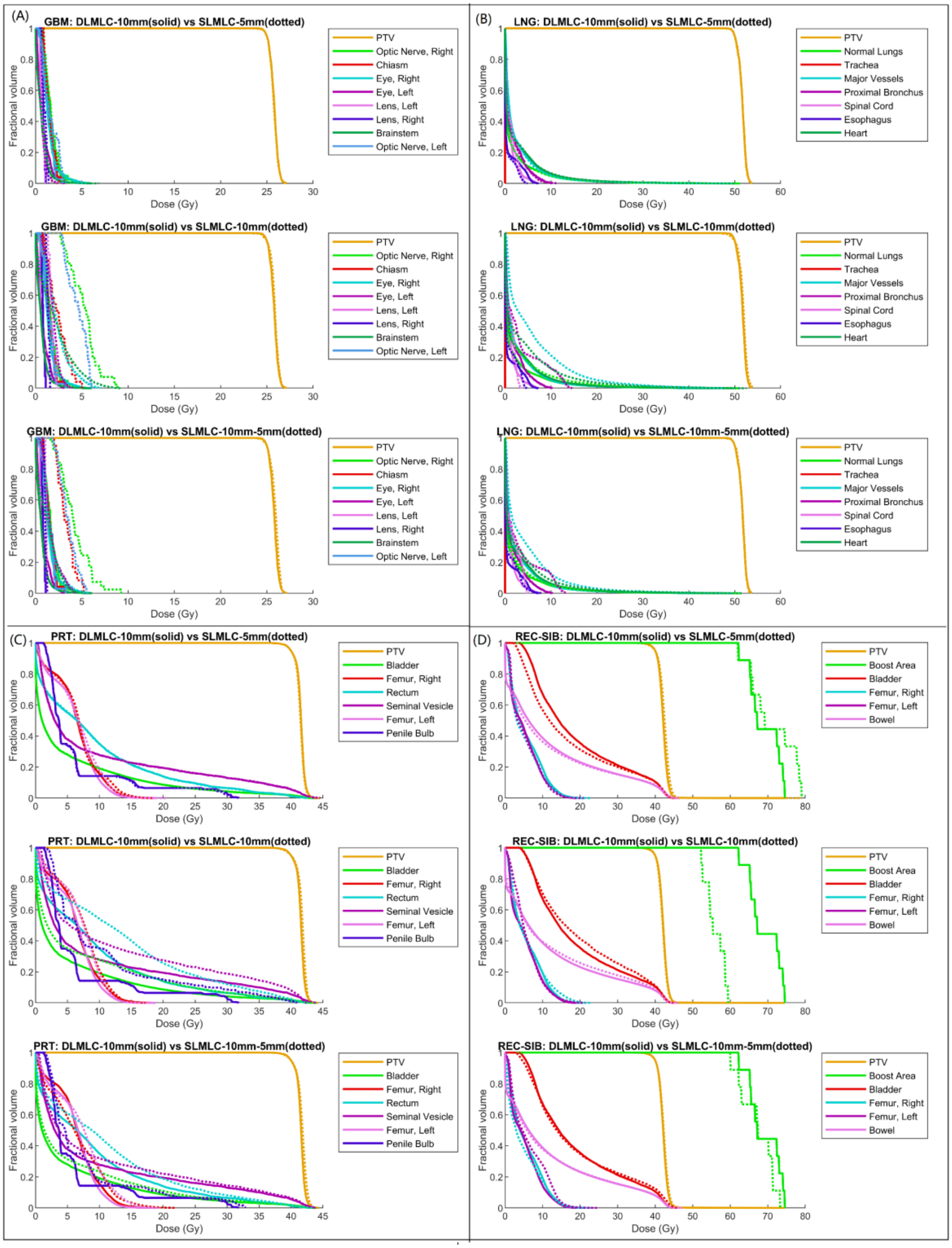


Figure 4-6 DVH for (A) the GBM case, (B) the LNG case, (C) the PRT case, and (D) the REC-SIB case. The solid lines are for the DLMLC plan, and the dotted lines are for SLMLC plans. D95 is normalized to the prescription dose.

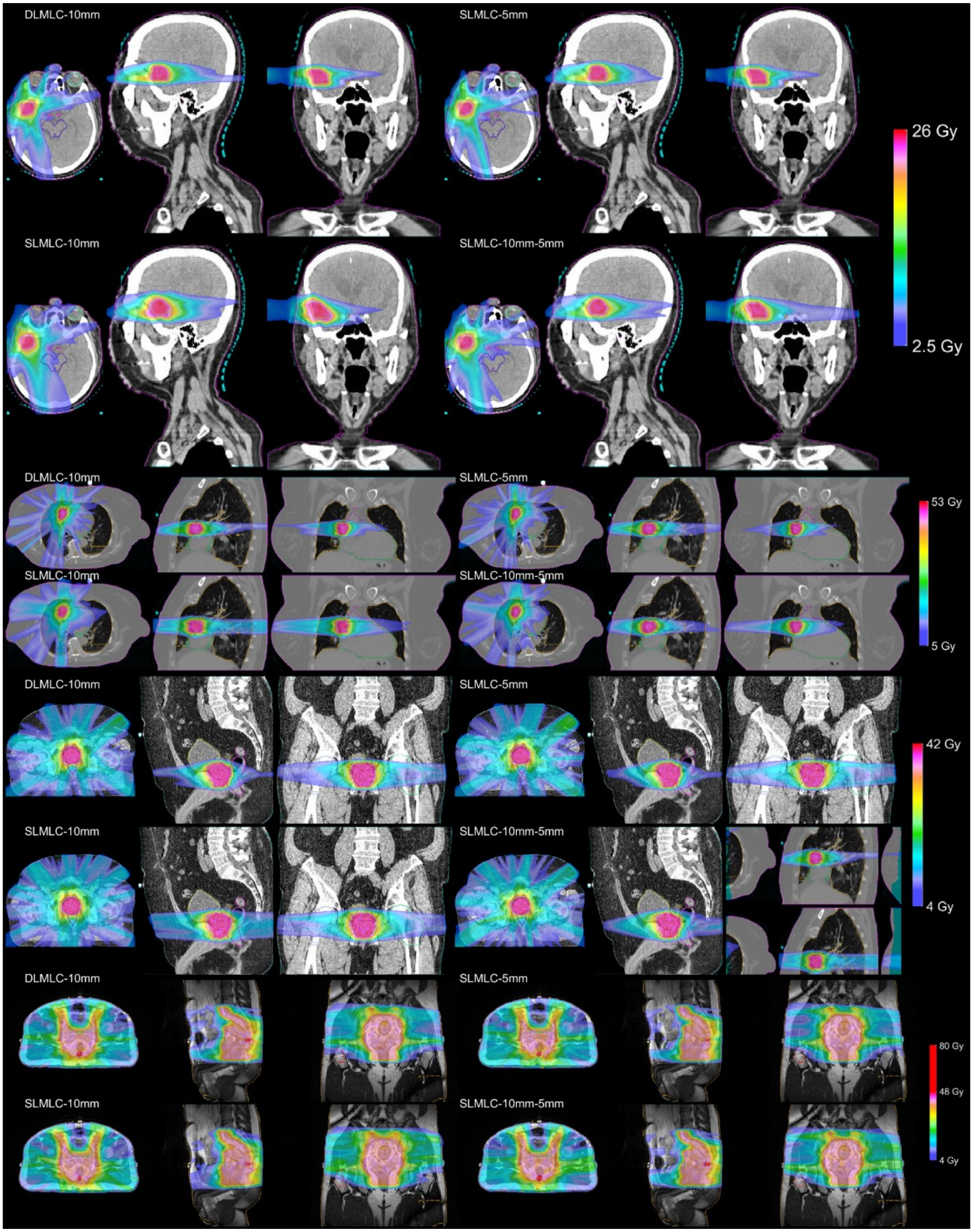


Figure 4-7 Isodose colorwash comparison for all patients

4.2.4 Discussion

Besides reducing the number of arcs, improvements of the linear accelerator in gantry speed, leaf speed, and dose rate have the potential to improve VMAT efficiency [22] further. This concept was recently implemented in a new ring gantry system (Varian Halcyon). The new system with enclosed gantry is designed for rapid treatment with increased gantry speed (4 rpm) and MLC leaf speed (5 cm/s) [102]. The higher mechanical performance has the potential to reduce treatment time and make treatment more robust to intrafractional motion. However, this potential is hindered by the current PSO optimization algorithm, which requires more arcs to achieve the same dosimetry as the systems with a single layer higher resolution MLC.

Our previous VMAT optimization framework for SLMLC [35] simultaneously optimizes 180 coplanar arc beams without progressive sampling. Apart from the optimization constraint, the previous SLMLC VMAT framework share similar objective function as the proposed DLMLC VMAT, including the least square dose fidelity objective, the anisotropic total variation term that regulates the fluence smoothness, and the single segment term for forming simple apertures. The previous SLMLC VMAT framework has been shown to outperform the PSO method that has been widely adopted in clinic [35], achieving improved dose conformality, reduced number of arcs, and shortened delivery time. In this study, we adapted the VMAT optimization framework for DLMLC by introducing novel optimization constraints. The adapted framework simultaneously solves for both MLC layers and 180 beams. With the same target coverage, the DLMLC-10mm plan substantially improved OAR sparing and dose compactness compared with SLMLC-10mm, achieving comparable plan quality with the SLMLC-5mm plan. This new optimization algorithm thus better takes advantage of the faster gantry rotation and leaf speed for more efficient plan delivery.

In each iteration, the DLMLC VMAT algorithm runs through 3 modules, optimizing with respect to the fluence map, the aperture shape, and the intensity value, respectively. This alternating optimization approach has been investigated on VMAT optimization in our previous studies [35,38,39]. Different from our previous studies, which used regularization terms to encourage deliverability, the current study formulates the deliverability constraint and leaf speed constraint as non-convex optimization constraints with respect to the aperture variable and then transforms it into a graph optimization problem, which is heuristically solved. This modification brought forth several benefits. First, the formulation maintains convexity with respect to the fluence map and the intensity value. Second, the number of hyperparameters are reduced compared with the original formulation, making parameter tuning easier. Third, the mechanical constraints with physical parameters are strictly enforced instead of being encouraged in the previous regularization approach. Last, the modification makes it straightforward to incorporate the DLMLC deliverability constraint, which is not obvious, if not impossible, to be formulated as a regularization term.

Compared with the previously reported Halcyon VMAT plans [98,102], which need more arcs to achieve comparable dosimetry to a VMAT plan on traditional TrueBeam C-arm linac, the proposed method clearly elevated the performance of single-arc DLMLC VMAT to be equal to SLMLC with higher resolution leaves. By overcoming the deficiency caused by optimization algorithms, DLMLC-10mm may replace the conventional 5mm SLMLC. This may apply to higher resolution MLCs, which have suffered from low fabricability and high inter-leaf leakage. As mentioned, although the Halcyon parameters are used in the current study, the proposed algorithm is in principle applicable to other DLMLC machines such as ViewRay with corresponding physical parameters.

Similar to the previous study in section 2.1, FISTA is adopted in this study for rapid optimization. The accelerated algorithm allows solving the full resolution DLMLC problem in 5 minutes for the GBM and LNG case, and 20 minutes for the PRT case. The SIB case took substantially longer time of 1 hour due to the large PTV size. For future clinical implementations, matrix multiplication, which is the most computationally expensive step in FISTA, can easily be parallelized using graphical processing units (GPUs) to improve computational efficiency [75].

In this study, the MLC transmission through a single layer of the DLMLC is ignored for simplicity. The single layer transmission can be included in the single segment term, such that it pushes the fluence map to be at the beam intensity for opened beamlets, at the transmitted intensity for the beamlets that are blocked by a single MLC layer, and 0 for the beamlets that are blocked by both MLC layers. Observable deterioration in dose happens when the transmission through a single layer of MLC is greater than 5%. In practice, due to the low transmission factor achieved by Halcyon (within 0.45% for single MLC layer [103]), this term did not make a perceivable difference in optimization results.

4.2.5 Summary

The novel VMAT optimization framework for DLMLC utilizes two MLC layers to afford more sophisticated modulation and improves the effective modulation resolution. Consequently, *single-arc* DLMLC VMAT achieves not only superior dosimetry to SLMLC VMAT with the same leaf width but also equivalent dosimetry to SLMLC VMAT of half the leaf width.

4.3 Many-isocenter Optimization for Robotic Radiotherapy.

4.3.1 Introduction

Since its initial introduction in 1982 [24], intensity-modulated radiotherapy (IMRT) has been compared favorably to the 3-dimensional (3D) conformal radiation therapy due to enhanced target coverage and improved organs at risk (OAR) sparing. Subsequently, the technologies involving inverse treatment optimization and multileaf collimators (MLC) for intensity modulation have been rapidly developed and adopted in the clinic. More recent technological advances such as helical TomoTherapy and volumetric modulated arc therapy (VMAT) further streamlined IMRT delivery while maintaining comparable plan quality [104–106]. Since further plan quality improvement is hampered by the limited beam geometry in the coplanar IMRT and VMAT mode, planning methods using optimized non-coplanar beam angles, termed 4π radiation therapy [107,108,117,118,109–116] have been developed. Compared with state-of-the-art coplanar IMRT methods, 4π significantly reduced high dose spillage to the normal tissue. The improved dose compactness is desirable for stereotactic radiotherapy where the normal tissue toxicities are manifested in the high dose region [119].

Despite the demonstrated dosimetric benefits, 4π radiotherapy clinical adoption is not straightforward. Studies have shown the feasibility of delivering 4π IMRT [120] and 4π VMAT [121,122] plans on the C-arm gantry platform, where the non-coplanar beam orientations require the combination of couch and gantry rotation to achieve. However, the combined motion increases treatment delivery time, risk of collision and unwanted patient secondary motion. Couch rotations also create challenges to maintain constant monitoring of the patient position. The more complex dry run and QA before treatment delivery is yet another obstacle.

Alternatively, a robotic linac platform is more conducive for the non-coplanar and non-isocentric treatment. For instance, the CyberKnife Robotic Radiosurgery System (Accuray, Sunnyvale, CA, USA) mounts a 6-MV linac to an articulated robotic arm that allows both non-coplanar and non-isocentric treatments without moving the patient. Currently, CyberKnife does not have access to all non-coplanar angles [123] but this engineering challenge may be overcome with a more compact linac design, modification to the robotic arm or the treatment room. Celestial Oncology (Santa Monica, CA, USA) is currently developing a more compact linac with the complete access to the 4π Steradian angles around the patient. However, to make the linac head smaller for the robotic platform, a more stringent requirement is placed on the MLC form factor. Both the number and the travel of the leaves are limited for the compact linac head size, creating a significant challenge in attaining both large field-of-view (FOV) and high modulation resolution. A method to achieve both goals would significantly increase the versatility of robotic linacs.

Thus far, 4π radiotherapy is mainly performed with the assumption of isocentric geometry and 100 cm source-to-isocenter distance (SID) [120,124], which is the native geometry of the ubiquitous C-arm gantry linacs with one degree of rotational freedom during treatment. Multiple isocenter treatments on a conventional C-arm linac is cumbersome and potentially hazardous due to the increased chance of geometric error and collision. It has been performed for large targets when the isocentric geometry cannot provide a sufficiently large FOV to cover the entire target, such as breast [125,126], supraclavicular fossa [126], craniospinal radiotherapy [127], and etc. [128,129].

From robotic linacs, varying the SID and adding the number and locations of isocenters as a new degree of freedom appears to be a logical extension of the 4π research due to the following considerations. 1. Using a smaller SID would allow a higher MLC modulation resolution that is dosimetrically beneficial in highly complex cases. 2. Using a smaller SID results in more rapid dose fall-off in the beam direction due to the more prominent inverse square effect. 3. The small single FOV can be compensated by using more than one isocenters to cover the entire tumor. 4. The treatment efficiency of using more isocenters is compensated by the increase in the dose rate at the shorter distance. On the other hand, most beam angles with substantially shorter distances, i.e., 50cm SID, are geometrically prohibited on the C-arm gantry.

In this study, we strategically divide a large target into many smaller targets each with its own isocenter, to achieve small FOV beam high-resolution delivery at a shorter distance. A many- isocenter planning problem was solved to simultaneously optimize the beams and fluence maps. In this study, to determine the dosimetric benefit of many-isocenter planning, we adopt the geometry of the new robotic linac that is currently under development at Celestial Oncology. However, methods used in this study are generalizable to any radiation delivery platforms that have access to large non-coplanar angles, non-isocentric beams, and substantially different source-to-tumor distances.

4.3.2 Methods

4.3.2.1 Robotic platform model

Figure 4-8(A) shows the robotic radiotherapy platform under development by Celestial Oncology, which reduced the linac head size by using a compact x-band 6MV source and reducing the distance between the X-ray source and the MLC. To study the dosimetry and delivery efficiency, the two SIDs are considered: 100 cm and 50 cm. At SID-100, the projected MLC leaf width is 1 cm, and the FOV is 20cm by 20cm with a total of 20 leaves per bank. The physical size and the number of the leaves were determined based on the fabrication practicality and size of the robotic linac head. At SID-50, the projected MLC leaf width is 0.5cm, and the FOV is 10cm by 10cm. For Head and Neck (H&N) cancer with the planning target volume (PTV) up to 20 cm, the target can be fully covered by a SID-100 beam (Figure 4-8(B)). For SID-50, beams of many isocenters are required to efficiently cover the entire target (Figure 4-8(C)).

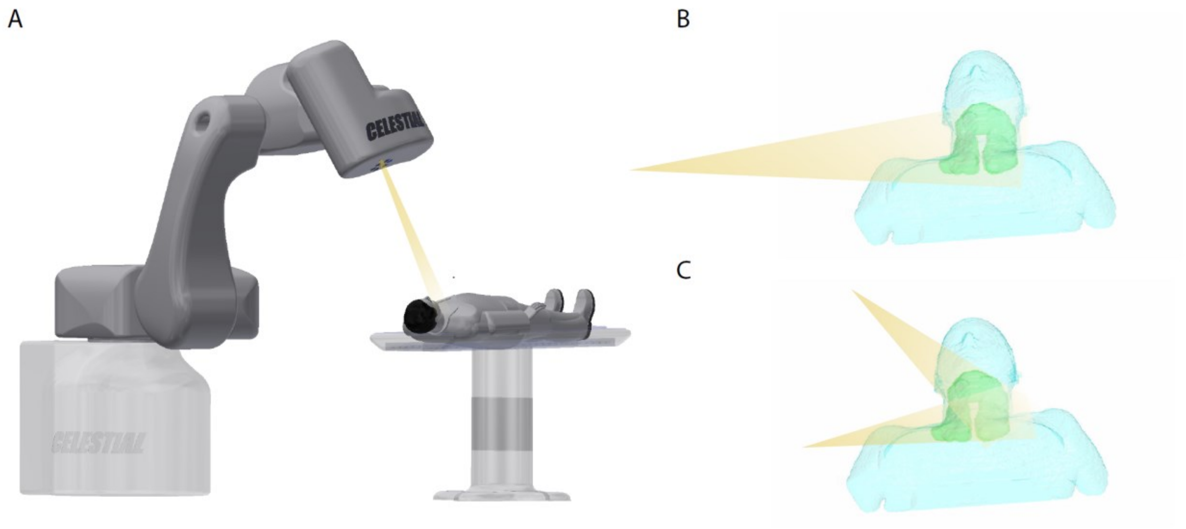


Figure 4-8 (A) Demonstration of the robotic arm platform, (B) an isocentric SID-100 beam that covers the entire target, (C) beams of different isocenters are required to efficiently cover the entire target.

4.3.2.2 Determine isocenter locations

We use the following method to determine the position of isocenters when the FOV is not large enough to cover the entire target. Figure 4-9 shows an illustration of the PTV, bounding box, and isocenters. The bounding box of a target is defined as the smallest cuboid that fully covers the entire PTV in the scanner coordinate system. The dimension of the bounding box is denoted by $[B_x, B_y, B_z]$, and the beam FOV is denoted by $[F_x, F_y, F_z]$. We first divide the bounding box to N_x, N_y, N_z number of identically sized boxes in the x, y, z directions respectively, where $N_w = \lceil B_w/F_w \rceil, w = \{x, y, z\}$. ($\lceil \cdot \rceil$ represents the ceiling operation). For each divided box, the isocenter position is set at the center of mass (CoM) of the partial PTV within the box. The rationale behind the method is that each isocenter will be 'in charge' of a partial PTV for which the x, y, and z dimensions are smaller than the nominal beam FOV. Note that this method locates a set of isocenters that are adequate to cover the entire PTV, but it does not necessarily demand all identified isocenters to be used in the final plan. The BOO algorithm, which is discussed in the section 4.3.2.3, determines whether the candidate beams and the isocenters are utilized in the final plan.

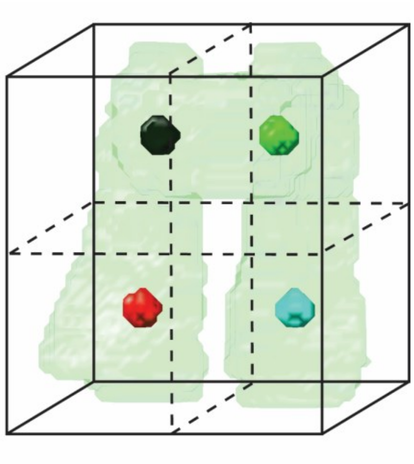


Figure 4-9 Illustration of the PTV, bounding box, and four isocenters.

4.3.2.3 Integrated Fluence Map Optimization, Beam Orientation Optimization, and isocenter selection.

The integrated framework for fluence map optimization (FMO), beam orientation optimization (BOO), and isocenter selection is formulated as

$$\underset{\boldsymbol{x}}{\textbf{minimize}} \quad \frac{1}{2}\|\boldsymbol{W}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{d})\|_2^2+\boldsymbol{\lambda}\|\boldsymbol{D}\boldsymbol{x}\|_1+\boldsymbol{\gamma}\sum_{\boldsymbol{b},\boldsymbol{i}}\boldsymbol{w}_{\boldsymbol{b},\boldsymbol{i}}\left\|\boldsymbol{x}_{\boldsymbol{b},\boldsymbol{i}}\right\|_2^{\frac{1}{2}}$$

$$\textbf{subject to} \qquad \boldsymbol{x} \geq \boldsymbol{0},$$

Equation 4-8

where x is the vectorized fluence map, A is the fluence to dose transformation matrix, d is the vectorized ideal dose distribution, with prescription dose at the PTV and 0 elsewhere. W is the diagonal structure weighting matrix. The matrix D is the derivative matrix, and λ is the corresponding weighting coefficient. $x_{b,i}$ is the vectorized fluence map of beam b and isocenter i, and $w_{b,i}$ is the beam weighting coefficient.

In this formulation, the data fidelity term attempts to find the optimal fluence map $\hat{x}$ such that the calculated dose is as close as possible to the ideal dose. The priorities for the structures of interest are controlled by the diagonal weighting matrix W. The second term is the total variation (TV) regularization which encourages piecewise continuity of the fluence map [130]. The amount of smoothness is controlled by λ.

The third term is the group sparsity term in the form of $l_{2,1/2}$ norm penalty, which encourages most candidate beams to be inactive. The individual beam weights w_b serve as a normalization to correct for the intrinsic norm differences among candidate beams, which has been widely used in compressed sensing [131]. The individual beam weights are calculated by $w_{b,i}=\left(\frac{\operatorname{mean}\left(A_{PTV}^{b,i}\vec{1}\right)}{\sqrt{n_{b,i}}}\right)^{1/2}$, where $n_{b,i}$ is the number of beamlets in beam b for isocenter i, and $A_{PTV}^{b,i}$ is the fluence to dose transformation matrix within the PTV, for beam b directed at isocenter i. The parameter γ controls the global sparsity level, which is tuned automatically to achieve the desired number of beams. At every 1000 iterations, the current number of selected beams is checked and compared with the desired number of beams. The value of γ is increased by a factor of 1.5 if there are more beams selected than required or decreased by a factor of 1.5 if there are not enough beams selected. The optimization problem was solved with FISTA [132].

4.3.2.4 Evaluation

The efficacy of the optimization algorithm was tested on 10 consecutive H&N patients who have been diagnosed with oropharyngeal cancer. The patients were CT-simed with immobilization mask. The contours and prescription follow 2018 American Society of Therapeutic Radiation Oncology (ASTRO) head and neck treatment guideline.

Patient	PTV volume	Bounding Box (cm)			Isocenter #		Sampled beam #		Candidate beam #	
		x	y	z	SID-100	SID-50	SID-100	SID-50	SID-100	SID-50
H&N #1	610.5	9.3	14.3	15.8	1	4	1162	4648	776	2785
H&N #2	724.6	10.8	16	18.3	1	8	1162	2320	826	1440
H&N #3	947	10	17	22.3	2	12	2324	3480	1579	2117
H&N #4	785	10.3	15.5	19	1	8	1162	2320	891	1452
H&N #5	686.4	10.5	14.5	18.3	1	8	1162	2320	824	1422
H&N #6	787	10.5	16.5	17.5	1	8	1162	2320	842	1443
H&N #7	352.7	8.5	8.5	18.5	1	2	1162	2324	758	1315
H&N #8	555.7	9.8	13.3	14.8	1	4	1162	4648	906	2904
H&N #9	271.3	8.8	6.8	17	1	2	1162	2324	781	1385
H&N #10	620.5	10.3	15	14	1	8	1162	2320	685	1278

Table 4-7 PTV volumes, PTV bounding box dimensions, number of isocenters, number of sampled beams, and number of candidate beams (non-colliding beams) for all patients.

Table 4-7 summarizes the PTV volumes, PTV bounding box dimensions, number of isocenters, number of sampled beams, and number of candidate beams (non-colliding beams) for all patients. For cases with fewer than or equal to four isocenters, for each isocenter, 1162 beams were uniformly sampled in the 4π space with 6° of separation. For cases with more than 4 isocenters, to reduce the computation cost of both dose calculation and optimization, for each isocenter 290 beams were uniformly sampled with 12° of separation, and the angular sampling of adjacent isocenters differ by 6° to compensate the coarse sampling. For example, if the angular sampling of one isocenter is $\{0°, 12°, 24°, \ldots\}$, then the angular sampling of the adjacent isocenter is $\{6°, 18°, 30°, \ldots\}$. A computer-aided-design (CAD) model of a robotic arm platform and a 3D optically reconstructed surface of a human subject were utilized to determine the patient-specific collision-free space [133]. After excluding the colliding beams, the number of candidate beams for each case can be found in Table 4-7.

The beamlet dose calculation used a convolution/superposition code with a 6 MV x-ray polyenergetic kernel, as described in our previous publication [134]. The beamlet resolution at the isocenter was 0.5×0.5 cm^2 for SID at 50 cm and 1×1 cm^2 for SID at 100 cm, adjusting for the beam divergence effect. The dose array resolution was $0.25\times0.25\times0.25$ cm^3.

As a sensitivity analysis, we compared the SID-50 plan quality against the SID-100 with the number of selected beams ranging from 15 to 50, based on the corresponding data fidelity values. The plans with 20 beams were used for further evaluation on plan efficiency and plan quality.

The delivery time was estimated for both a hypofractionation scenario (assuming 5 fractions), and a conventional fractionation scenario (assuming 30 fractions), for all 20-beam plans. The estimation includes the times for MLC leaf travel, x-ray delivery, and Linac head travel. To estimate the MLC leaf travel time, the fluence map was first stratified into a finite number of discrete levels, and then sequenced using a reducing level method [135]. Each sequenced segment was further divided into a few deliverable segments, and all segments within one beam were reordered to reduce MLC leaf motion by solving a traveling salesman problem (TSP) [136]. The same algorithm was used to reorder the beams and minimize the amount of Linac head travel. Note that the stratification was performed only for evaluating delivery efficiency, and all dosimetric evaluations were based on the theoretical dose from optimization.

In the efficiency estimation, the following assumptions on the mechanical specifications were used:

1. The MLC travel speed is at 2.5cm/s at 100cm from the source, and 1.25cm/s at 50cm.
2. The Linac head rotation speed is 1 rpm with respect to the isocenter according to the International Electrotechnical Commission (IEC) safety standard.
3. The effective dose rate is 500MU/min at 100cm from the source accounting for small field output factor and flattening-filter-free dose heterogeneity, and 2000MU/min at 50cm from the source. (The maximum dose rate of the robotic linac is 1000MU/min at 100cm from the source in the flattening filter free mode.)

PTV statistics including PTV D95, D98, D99, maximum dose (Dmax), and PTV homogeneity (defined as $\frac{\mathrm{D95}}{\mathrm{D5}}$) were evaluated. The maximum dose is defined as D2 (the dose at 2% of the structure volume), following the ICRU-83 report [101]. The dose conformity, R50, and integral dose were also assessed to quantify the dose conformity, compactness, and total spillage, respectively. The dose conformity is defined as the ratio between the patient volume receiving 100% or more of the prescription dose and the PTV volume. The R50 is defined as the 50% isodose volume divided by the target volume. For OAR, the Dmax and mean dose (Dmean) were obtained. Wilcoxon signed-rank test was conducted to determine whether a significant difference exists between the SID-50 and the SID-100 for all PTV and OAR statistics.

4.3.3 Results

Figure 4-10 shows the final dose fidelity values of SID-100 and SID-50 plans, varying the number of selected beams between 15 and 50. The average optimization runtime is 25 minutes for 15-beam plans, and 22 minutes for 50-beam plans, on an intel Xeon E5-2670 CPU with 8 physical cores and a base clock speed of 2.6 GHz. For each patient, the dose fidelity values were normalized by the average dose fidelity values of all plans associated with that patient. The plot with error bars shows a summary of all patients. Overall, SID-50 plans achieved lower dose fidelity value compared with the SID-100 plans using the same number of beams, showing superior plan quality to the SID-100 plans. The gap between SID-50 and SID-100 widens with an increasing number of beams.

Each patient plot is titled with the patient number, the number of isocenters for the SID-50 plan, and the number of isocenters for the SID-100 plan. For example, the first patient plot is entitled: '#1: 4(50), 1(100)', showing that patient #1 has four isocenters for the SID-50 plan and one isocenter for the SID-100 plan. For small tumors with only two isocenters (#7, #9) or four isocenters (#1, #8) in the SID-50 plans, SID-50 achieves an unquestionable advantage among all plans. For medium-sized tumors that require 8 isocenters in the SID-50 plans (#2, #4, #5, #6, #10), SID-50 shows clear advantage with 20 or more beams. For the large tumor that requires 12 isocenters (#3) in SID-50 and two isocenters in SID-100, the SID-50 is comparable to SID-100 with fewer than 20 beams, but the SID-50 plan is increasingly better with more than 20 beams.

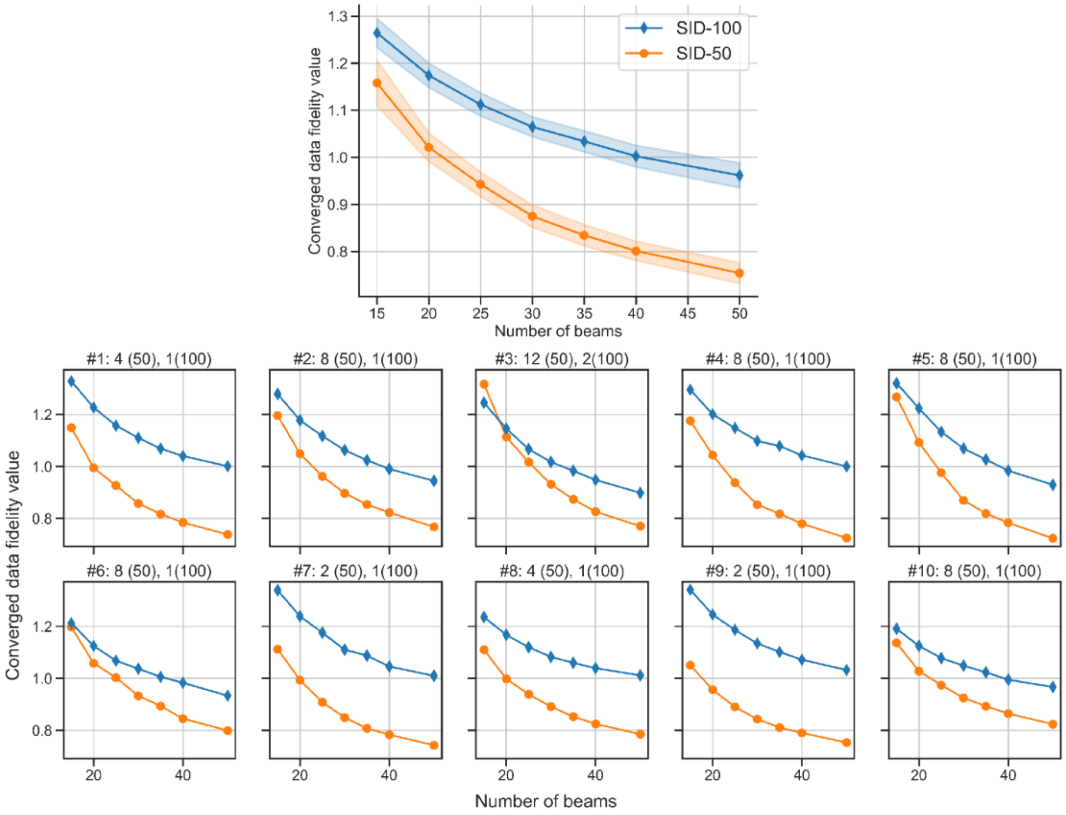


Figure 4-10 Final data fidelity value vs the number of beams. The plot with shaded error bar shows a summary of all patients. Each patient plot is titled with the patient number, the number of isocenters for the SID-50 plan, and the number of isocenters for the SID-100 plan. For example, the first patient plot is entitled: '#1: 4(50), 1(100)', showing that the patient #1 has four isocenters for the SID-50 plan, and one isocenter for the SID-100 plan.

Figure 4-11 shows the estimated delivery time for all 20-beam plans assuming (A) 5 fractions and (B) 30 fractions for all patients. The delivery time includes three modules: MLC travel time, MU delivery time, and linac head travel time, among which the latter takes the least amount of time, with only around 4s per beam. On average, for a conventional fractionated treatment plan (30 fractions), it takes 14 minutes to deliver a SID-50 plan, and 9 minutes to deliver a SID-100 plan. The longer delivery time of SID-50 is attributed to a longer MLC travel time, which is the most time-consuming module in a 30-fraction treatment. Due to the higher modulation capability of SID-50, the corresponding fluence map is more complicated, requiring more MLC segments in each beam and a longer MLC leaf travel time. For hypofractionation (or Stereotactic Body Radiation Therapy (SBRT)), SID-50 delivery takes 19 minutes, and SID-100 delivery takes 18 minutes. In SBRT, each fraction delivers a greater dose, and the beam-on time is more dominant. The dose rate of SID-50 is four times that of SID-100 due to the inverse square effect. Therefore the 'beam-on' time of SID-50 is substantially shorter. The total delivery time of SID-50 and SID-100 are comparable for SBRT.

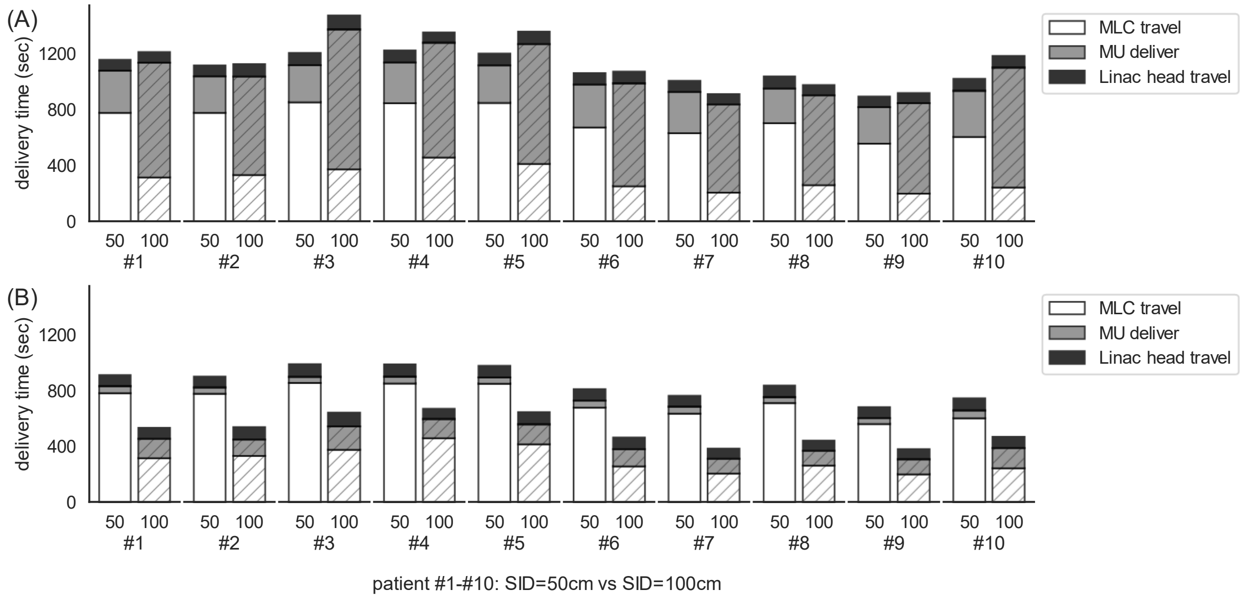


Figure 4-11 Delivery time estimation assuming (A) 5 fractions and (B) 30 fractions for all 20-beam plans.

Figure 4-12 shows the selected beams in the 20-beam plans of all patients. Note that not all isocenters were utilized in the final plan. For example, for patient #5, the isocenter on the upper, right, and posterior side of the tumor is not utilized. This is partly due to that there are enough beams coming from the left side of the patient, and that the partial PTV of the isocenter has a relatively lower prescription dose. The omission of certain isocenters for efficient delivery is one of the advantages of the integrated optimization framework. Such a decision would be unattainable by a human planner or a greedy algorithm based on only the patient anatomy.

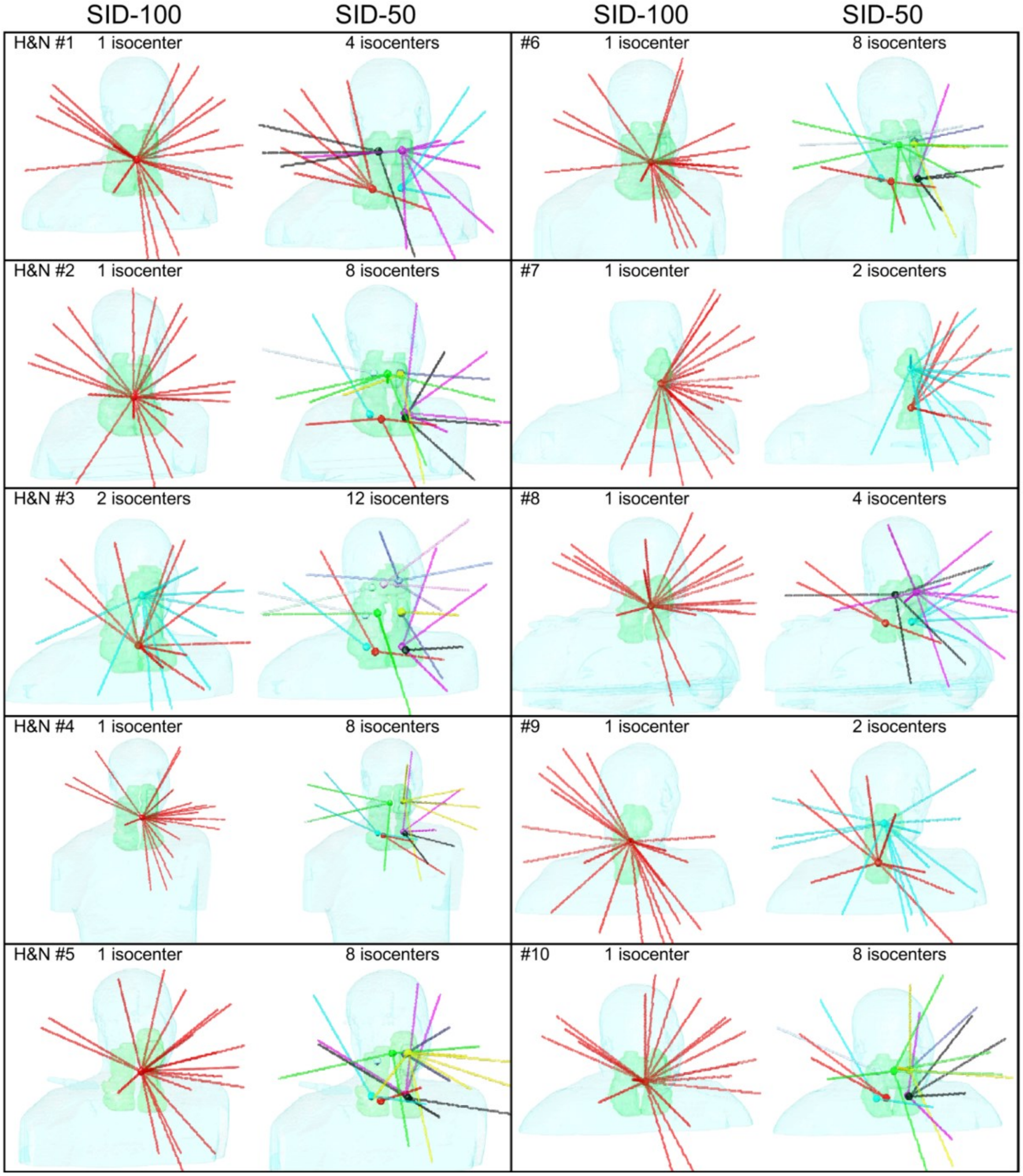


Figure 4-12 Selected beams in the 20-beam plans of all patients.

Figure 4-13 shows the isodose colorwash of SID-100 and SID-50 for two patients (#1 and #2, having 4 and 8 isocenters in the SID-50 plan, respectively). Overall, SID-50 achieved a more compact dose distribution. For patient #1, as indicated by the red arrows on the transverse view and coronal view, the SID-50 plan formed a more pronounced dose valley between the two separate PTVs, demonstrating the advantage of higher modulation resolution. In addition, the dose to the spinal cord and mandible was substantially reduced compared with SID-100. For patient #2, the isodose shows reduced low dose spillage in the brain and reduced high dose spillage to the nearby critical organs such as the spinal cord and trachea.

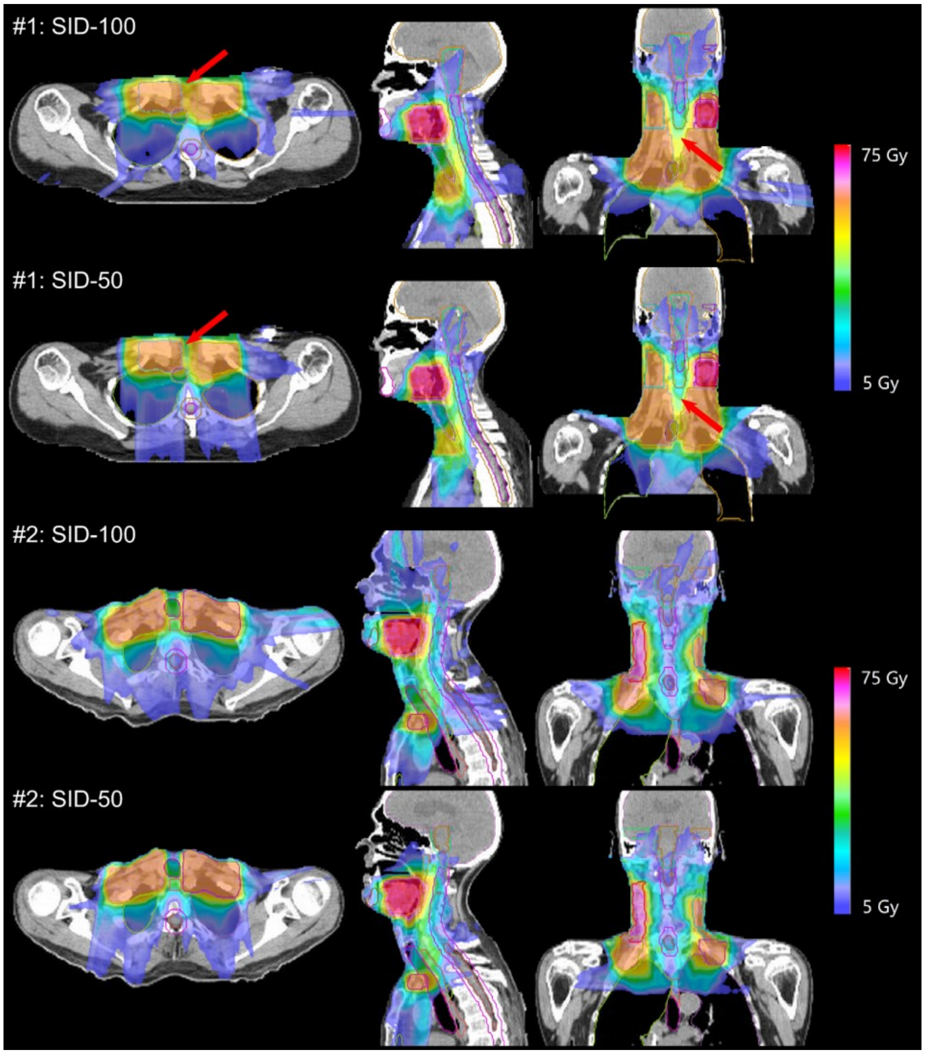


Figure 4-13 Isodose colorwash of 20-beam SID-100 and SID-50 plans for two patients (#1 and #2, having 4 and 8 isocenters in the SID-50 plan respectively).

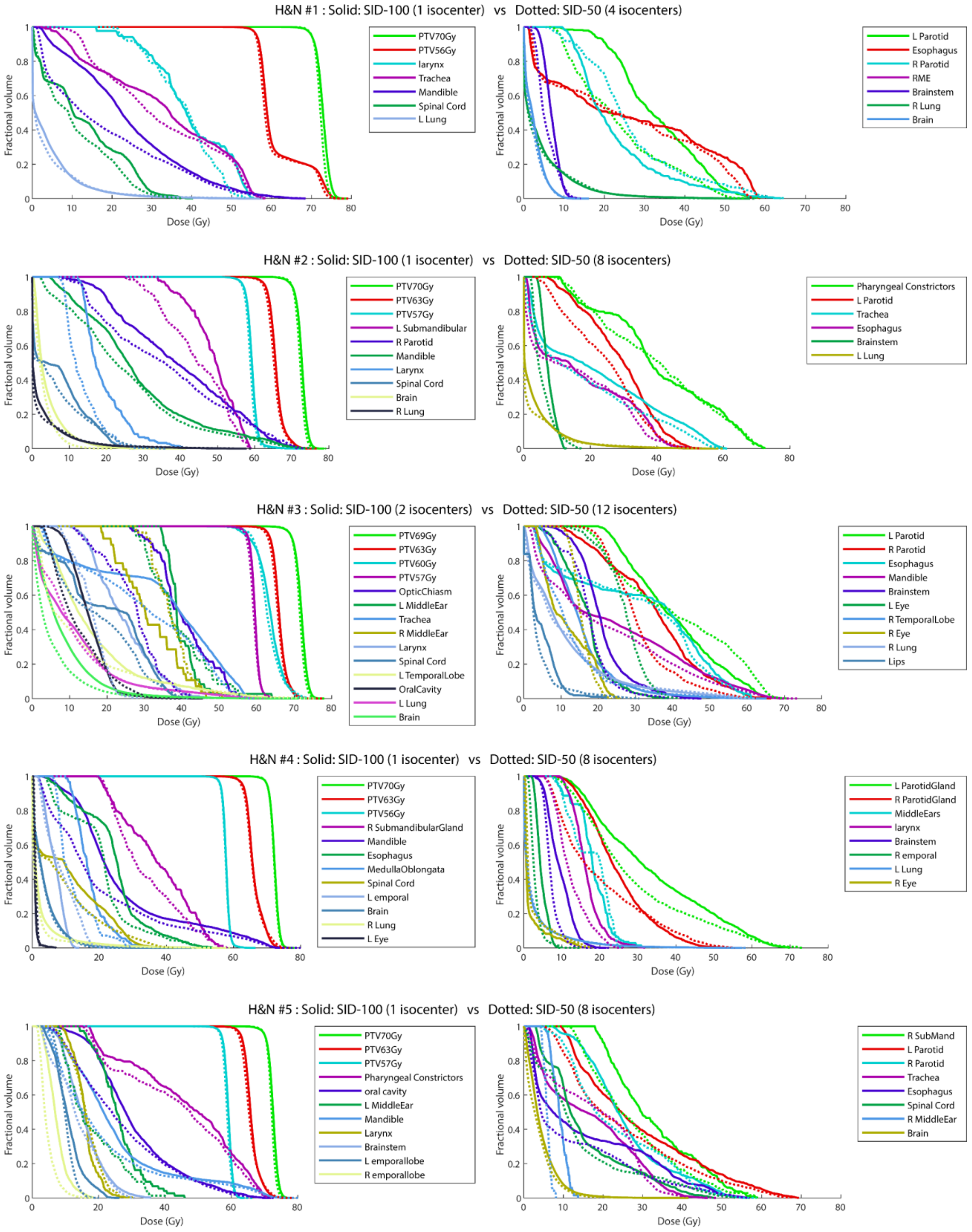


Figure 4-14 DVH comparison of the 20-beam SID-100 plans (solid) and SID-50 plans (dotted) for patient #1 - #5.

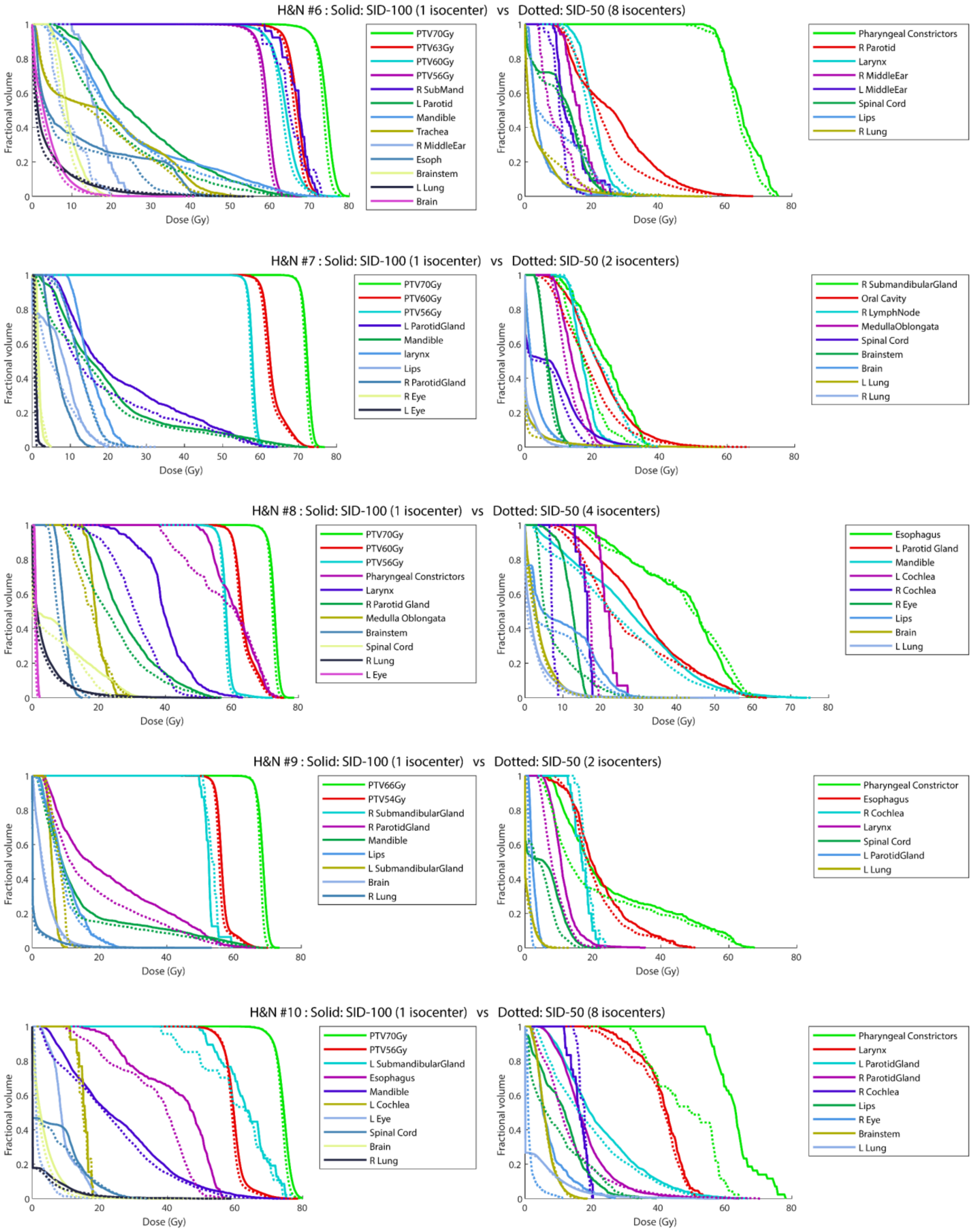


Figure 4-15 DVH comparison of the 20-beam SID-100 plans (solid) and SID-50 plans (dotted) for patient #6 - #10.

Figure 4-14 and Figure 4-15 show the DVH comparison of the 20-beam SID-100 plans (solid) and SID-50 plans (dotted). SID-50 improved PTV coverage, reduced hot spots within PTV, and reduced mean and max dose for most OARs. For patient #1, the dose to the left parotid, mandible, and spinal cord was substantially reduced in the SID-50 plan. For patient #2, SID-50 markedly improved sparing of the dose-limiting organs, such as the left submandibular gland, right parotid gland, left parotid gland, and trachea.

Figure 4-16 shows the PTV statistics, including homogeneity, D99, D98, D95, and D2, for all PTVs (29 PTVs in total). The same PTV in the SID-50 plan (orange) and the SID-100 plan (blue) are connected with lines. Overall, SID-50 achieved comparable PTV D95, D98, D99, improved PTV homogeneity, and reduced hot spot within the target, indicated by the lower D2 (maximum dose) values.

Figure 4-17 shows the dose conformity, R50, and integral dose of all plans using 20 beams. The dose conformity was comparable between SID-50 and SID-100, but SID-50 reduced R50 by 5.3% and integral dose by 9.6%, indicating a remarkable improvement in dose compactness and overall low dose spillage.

Figure 4-18 shows the mean and maximum dose for OAR overall and selected critical OARs including parotid gland, larynx, and mandible, for all plans using 20 beams. On average, the SID-50 plan reduced the [Dmean, Dmax] by [2.09 Gy, 1.19 Gy] for OAR overall, [3.05 Gy, 0.04 Gy] for parotid gland, [3.62 Gy, 5.19 Gy] for larynx, [3.27 Gy, 1.10 Gy] for mandible.

Table 4-8 reports the Wilcoxon signed rank test of SID-100 and SID-50 for all PTV and OAR statistics. SID-100 and SID-50 are significantly different ($p<0.05$) in PTV homogeneity, PTV maximum dose (D2), R50, Integral dose, OAR max and mean dose. For example, the Wilcoxon signed rank test indicated that the R50 of SID-50 (M=2.254; SD=0.256) was statistically significantly lower than the R50 of SID-100 (M=2.381; SD=0.236), $p < 0.05$.

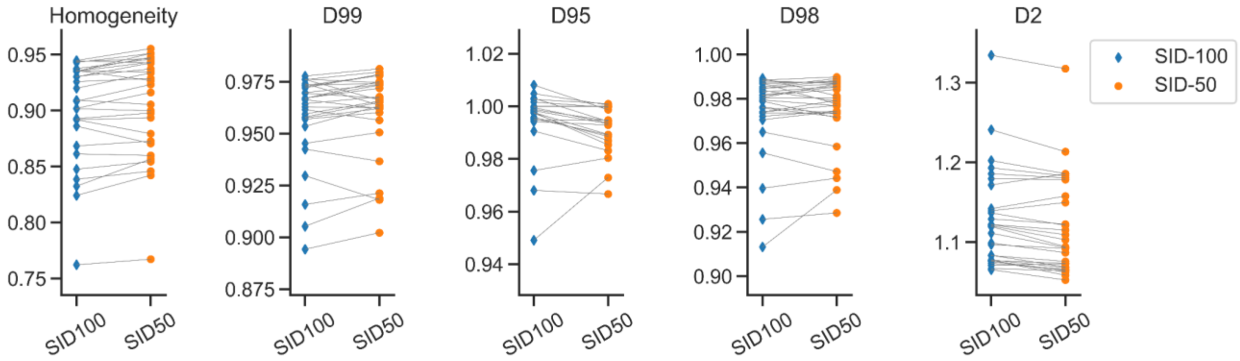


Figure 4-16 PTV statistics comparison of the 20-beam SID-100 (blue) and SID-50 (orange), for all PTVs (29 PTVs in total). The same PTV of the two plans are connected with lines. D2, D95, D98, and D99 are normalized by the prescription dose.

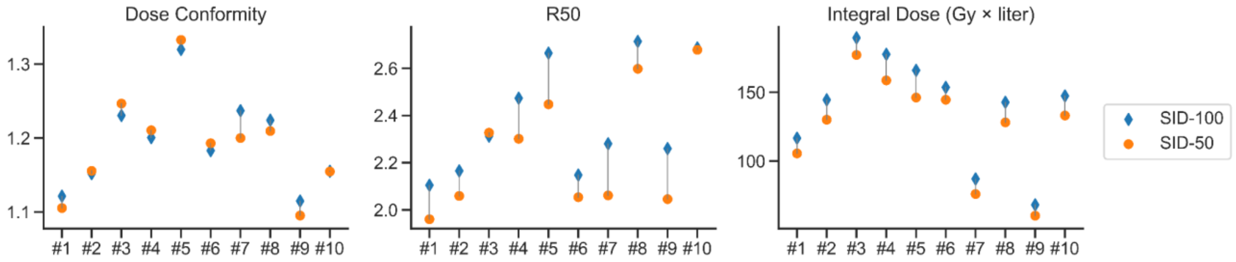


Figure 4-17 Dose conformity, R50, and integral dose for all patients (#1-#10). All plans have 20 beams in total. The SID-100 plan (blue) and the SID-50 plan (orange) of the same patient are connected with lines.

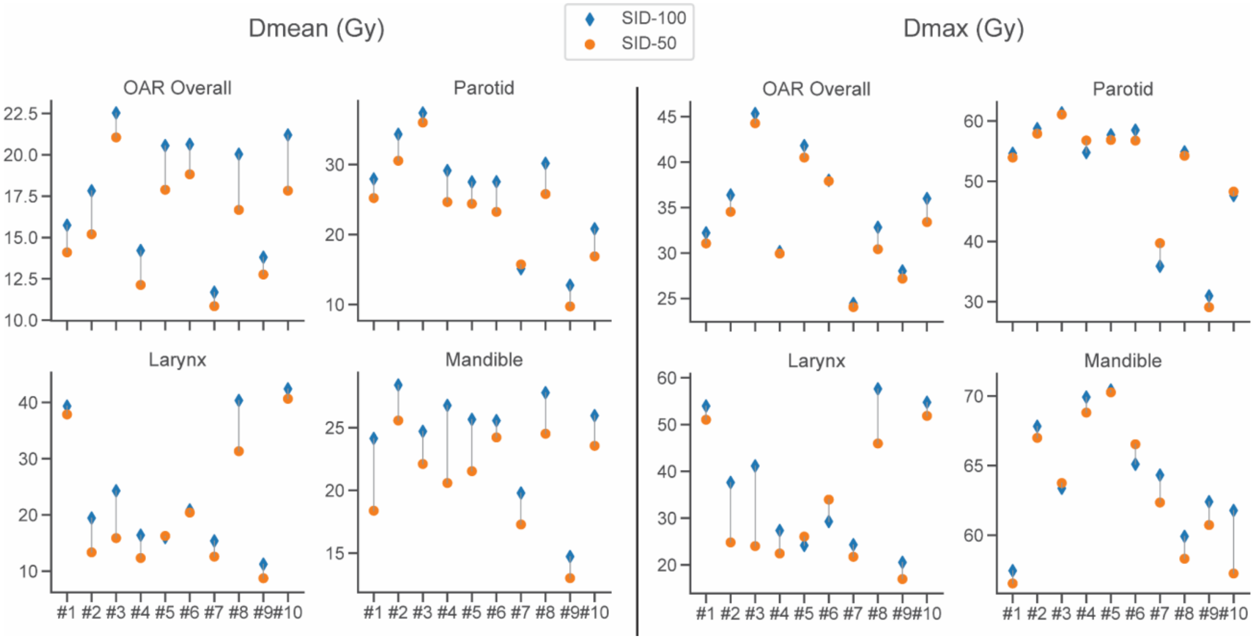


Figure 4-18 Mean and maximum dose for OAR overall and selected critical OARs including parotid gland, larynx, and mandible, for all patients (#1-#10). All plans have 20 beams in total. The SID-100 plan (blue) and the SID-50 plan (orange) of the same patient are connected with lines.

Statistics	p-value	Signed rank	Number of samples	SID-100		SID-50		(SID-50) - (SID-100)	
				Mean	STD	Mean	STD	Mean	STD
PTV Homogeneity	0.003	85	29	0.901	0.045	0.906	0.045	0.005	0.009
PTV D99	0.062	131	29	0.958	0.022	0.96	0.021	0.002	0.006
PTV D98	0.749	202	29	0.974	0.019	0.975	0.016	0.001	0.007
PTV D2	<.001	380	29	1.127	0.062	1.118	0.061	-0.009	0.01
Dose Conformity	0.625	33	10	1.194	0.062	1.19	0.069	-0.003	0.017
R50	0.006	53	10	2.381	0.236	2.254	0.256	-0.127	0.083
Integral Dose	0.002	55	10	139.377	38.481	126.02	36.139	-13.357	3.911
OAR Dmean	0.002	55	10	17.822	3.727	15.729	3.273	-2.094	0.894
OAR Dmax	0.002	55	10	34.515	6.317	33.325	6.173	-1.19	0.874

Table 4-8 Wilcoxon signed rank test of SID-100 and SID-50 for the PTV and OAR statistics in the 20-beam plans.

To demonstrate the benefits of using non-coplanar beams, we also compared the 4π plans with the coplanar plans for patient #1 and #2. Figure 4-19 show the DVH comparison of the 20-beam 4π IMRT plans (solid) and 20-beam SID-100 coplanar IMRT plans (dotted) for patient #1 and #2. With similar PTV statistics, 4π plans substantially reduced the dose to OARs. Figure 4-20 show the DVH comparison of the 20-beam 4π IMRT plans before (solid) and after (dotted) the stratification and sequencing, for both SID-50 and SID-100 plans of patient #1. The stratification and sequencing had minimal effect on the dosimetry with the integrated optimization approach.

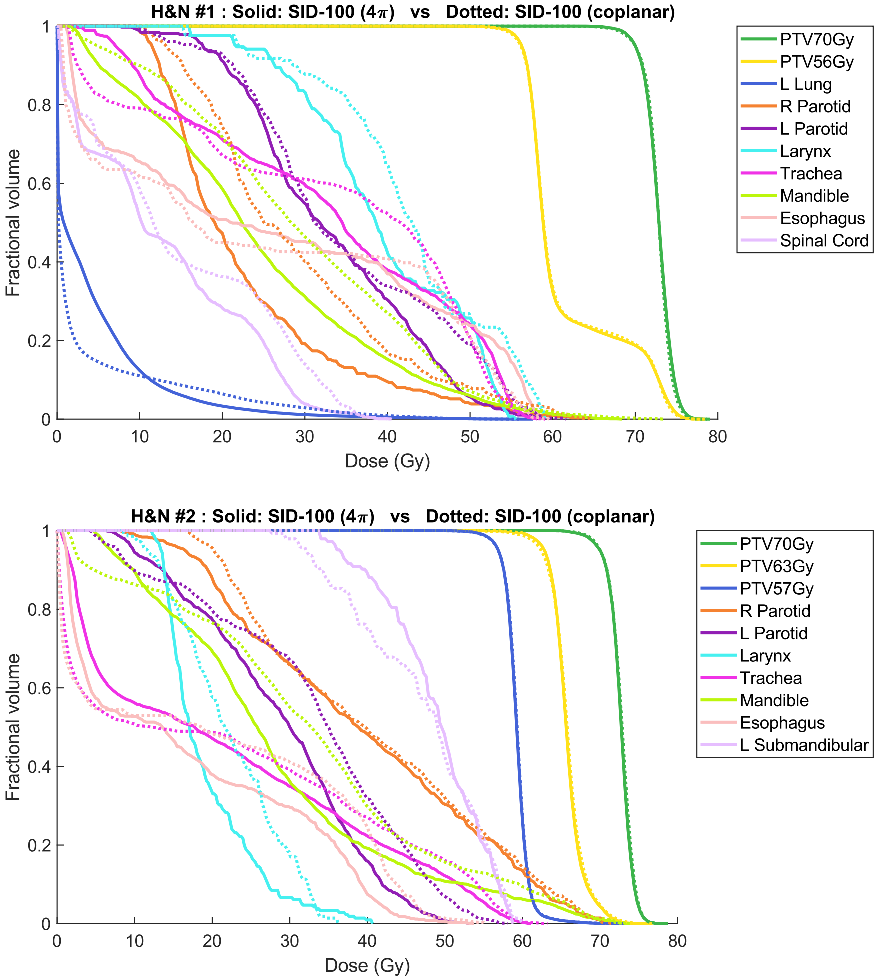


Figure 4-19 DVH comparison of the 20-beam SID-100 4π IMRT plans (solid) and 20-beam SID-100 coplanar IMRT plans (dotted) for patient #1 and #2.

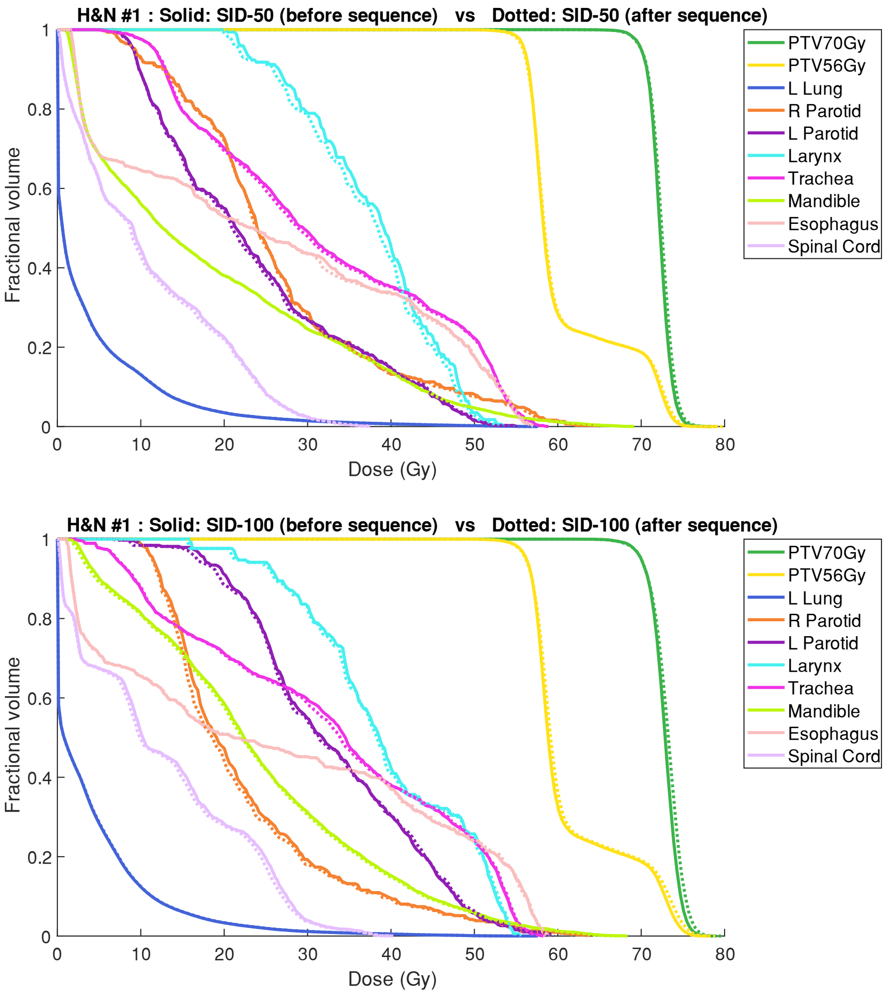


Figure 4-20 DVH comparison of the 20-beam 4π IMRT plans before (solid) and after (dotted) the stratification and sequencing, for both SID-50 and SID-100 plans of patient #1.

4.3.4 Discussion

Recent progress in non-coplanar radiotherapy reveals significant dosimetric improvement from the coplanar IMRT and VMAT [39,108,110–114,137–139]. But the clinical translation is hindered by the difficulty of delivering non-coplanar beams on the conventional C-arm platform. CyberKnife, which mounts the linac head on an articulated robotic arm, was expected to improve the dosimetry substantially with extensive usage of the non-coplanar space. However, besides the unclear dosimetric benefit and low treatment delivery efficiency due to the lack of posterior beams and the limitation from its heuristic optimization algorithm

[140], CyberKnife is limited in its versatility to treat large tumors. Therefore, the new robotic linac platform currently under development must overcome the following three inter-correlated challenges: 1. A more effective inverse optimization algorithm to create a superior quality treatment plan that can be efficiently delivered. 2. More compact linac head and modified robotic arm to access all 4π steradian angles. 3. The ability to treat large tumors without sacrificing efficiency and modulation resolution. Our previously developed optimization algorithm for 4π IMRT on the C-arm platform paved the foundation to solve the first challenge[141]. The second challenge can be overcome by the development of a more compact 6MV linac and using flexible distances between the source and the MLC. Here, we describe a solution to the third problem, which is to solve an integrated many-isocenter planning problem. The term “many-isocenter” is used based on two considerations. First, it provides the necessary differentiation from conventional “multiple-isocenter” treatment planning without integrated isocenter selection, beam orientation, and fluence map optimization. Second, the platform affords a significant increase in the isocenter number. Twelve isocenters were used in one case but more is possible, adding a new degree of freedom to static beam IMRT and non-coplanar modulated arc therapy [39] optimization. In the current implementation, the new degree of freedom affords both the large tumor coverage and high-intensity modulation resolution. With the same number of beams, the many-isocenter plans significantly improved the dose compactness as indicated by R50 and reduced both OAR and integral doses, thereby achieving overall superior plan quality to the single isocenter plans delivered at twice the distance.

The plan efficiency comparison depends on the type of treatment. For regular fractionated radiotherapy, 20-beam SID-50 plans are on average 36% more time-consuming than the SID-

100 plans with the same number of beams. For SBRT patients, assuming both plans have the same number of beams, the delivery time of SID-50 and SID-100 plan is comparable, and SID-50 has evident dosimetric advantages. As shown by Figure 4-10, there is space for further dosimetric improvement by allowing more beams if the dosimetry is prioritized over the delivery efficiency. Therefore, our method offers a solution to a basic conflict between the field-of-view and the intensity modulation resolution in robotic radiotherapy, and to an extent, C-arm gantry radiotherapy. By resolving this conflict, both high MLC resolution and large FOV can be achieved without sacrificing delivery efficiency.

In the past, numerous studies have reported stochastic and heuristic algorithms for BOO in IMRT, such as genetic algorithms [142,143], simulated annealing [144], and column generation [145]. Among them, the column generation algorithm has been extensively used for recent 4π radiotherapy research [114,124,146–148]. Column generation is a greedy method that iteratively adds new beams to the beam pool based on previously selected beams. A practical limitation is that the runtimes do not scale well with the number of beams to be selected, because it requires solving a larger optimization subproblem as more beams are added to the selected beam pool [141]. Although coplanar IMRT only requires 7 to 9 beams, it has been shown that the use of more beams is desirable for non-coplanar delivery (Dong et al., 2013), and possibly, even more, should be used for non-coplanar treatment with many isocenters. In this study, we used a group-sparsity regularized beam selection algorithm, which starts with all candidate beams and gradually reduces the number of active beams until the desired number is reached. This non-greedy approach is insensitive to the desired number of beams.

Due to the computation challenge of dose calculation/optimization of the vast combination of all possible beams, this study is restricted to SID of 50cm and 100cm, and the number of isocenter candidates is heuristically determined. In a more general many-isocenter optimization scenario, both SID and location of the isocenters could be optimized for greater dosimetric and delivery efficiency versatility. New mathematical paradigms may be needed to solve the optimization problem that is orders of magnitude larger.

4.3.5 Summary

This study investigates the feasibility of 4π radiotherapy using many isocenters on a robotic platform to treat large targets with reduced SID, which effectively increased the modulation resolution. Without sacrificing delivery efficiency, the many-isocenter plans delivered at shorter SID significantly improved dose compactness and OAR sparing compared with large FOV plans delivered at a longer SID. The group sparsity-based beam selection algorithm is applicable to the following two studies in section 4.4 and 4.5 to optimize collimator angles and gantry/couch angles, respectively.

4.4 VMAT optimization with dynamic collimator rotation

4.4.1 Introduction

In the current VMAT implementation, the collimator angle is kept static in each arc[22,149], which we term static collimator VMAT (SC-VMAT) for clarity. The large-scale VMAT optimization problem is solved by progressively increasing the beam angle sampling resolution until the full resolution, typically 2-degree separation between beams, is achieved. This progressive sampling approach is computationally straightforward to reach a good balance between optimization speed and plan deliverability, generating deliverable plans that meet the mechanical constraints of the multi-leaf collimator (MLC). On the other hand, there are several intrinsic limitations with this method. First, the greedy progressive sampling method is heavily influenced by the initial coarse resolution optimization results and can be trapped in an undesirable local minimum solution. Second, the interpolation between adjacent beams can lead to undesired struggles when there is a large change in the target projection shape with gantry rotation[94]. Third, optimization parameter tuning can be tricky as the weightings in different resolution stages of the optimization have a very different impact on the plan quality. Since the plan weightings in various stages of optimization are not easily traceable, exact reproduction of a plan can be difficult, even given the final plan optimization parameters. Last, despite its original promise of single arc optimization, two or more arcs are often used to improve the optimization results. To overcome these limitations, we developed a level-set based direct aperture optimization (DAO) for SC-VMAT[35]. In this approach, the fluence maps of individual beam directions are segmented using its level set function, and the entire VMAT optimization problem is efficiently solved in full angular resolution without downsampling. The efficient solution is enabled by a novel hybrid proximal primal-dual algorithm. The new single arc VMAT algorithm was shown to outperform the original progressive sampling VMAT method using two arcs. An important motivation of the original two-arc VMAT method is the dosimetric benefit of collimator rotation between arcs. By rotating the collimator a typically 90°, the MLC moving directions are orthogonal to each other between the two arcs. Although VMAT optimization is performed based on symmetric square beamlets, the two sides of the square are defined differently by the MLC leaf width and traveling motion, respectively, and are effectively asymmetric. The orthogonal arrangement of the two-arc VMAT MLC is able to mitigate the asymmetry and improve the dose distribution[150]. However, in this conventional approach, the collimator rotation is used as a user-selected constant instead of an optimization parameter that allows greater freedom to choose more than two collimator rotation angles. Studies[151] have shown an improved dose profile by using an adjusted collimator angle at each section of the arc in a full-arc VMAT with 4 sections, as compared with a full-arc SC-VMAT plan. In a heuristic approach to utilize the freedom of collimator rotation within the arc, Zhang et al.[152] developed a collimator trajectory selection paradigm for VMAT in paraspinal SBRT that aligns the collimator angle with the primary cord orientation. This study shows that by simply aligning the collimator with the target shape, a better dose distribution can be achieved. More recently, a beam's-eye-view approach was also proposed to guide the aperture selection for rotational delivery[153]. However, for sites with complex target and OAR sparing requirements, the geometrical approach is inadequate. Instead, a dose domain optimization approach is necessary to select the optimal collimator angle for each gantry angle while still maintaining the mechanical feasibility. In this study, we introduce a novel way of incorporating the dynamic collimator rotation into the single arc non-progressive sampling VMAT framework for dynamic collimator rotation enabled VMAT (DC-VMAT).

4.4.2 Methods

A flowchart of the proposed DC-VMAT optimization algorithm is shown in Figure 4-21. The optimization alternates between DAO and collimator angle selection (CAS). The DAO includes 3 modules, which solves the DAO optimization problem with respect to the fluence map $f_{b\alpha}$, the fluence intensity $c_{b\alpha}$ within an aperture, and the aperture variable $u_{b\alpha}$

respectively. Each iteration of the DAO runs through the 3 modules and optimizes one optimization variable, while holding the other two variables constant. This process is repeated until convergence. The CAS is performed on a simplified Dijkstra's map[154], generated from the result of DAO. The selected collimator angle $P_{b\alpha}$ is translated into the DAO in the next iteration, penalizing unselected candidate beams and encouraging development of the fluence map for selected beams. Formulations of DAO and CAS are presented in sections 4.4.2.1 and 4.4.2.2, and the corresponding algorithms are discussed in section 4.4.2.3.

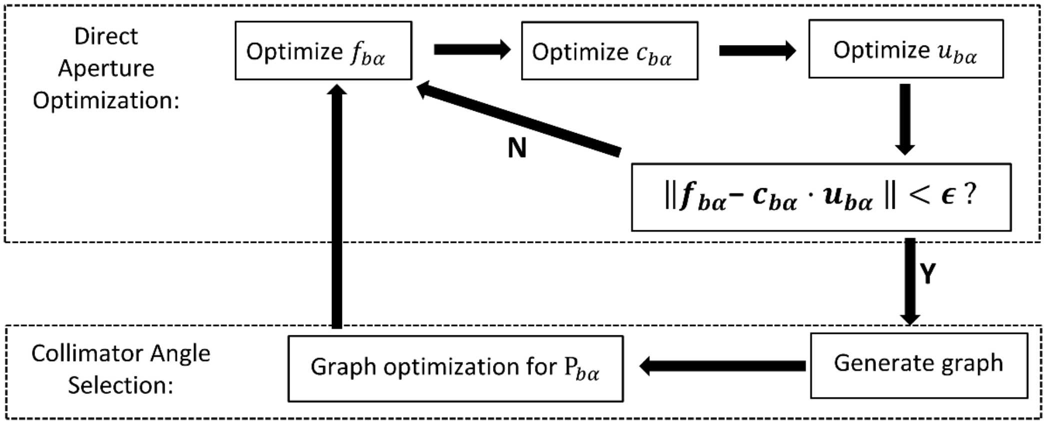


Figure 4-21 Flowchart of DC-VMAT optimization

4.4.2.1 <u>Direct Aperture Optimization</u>

The proposed DC_VMAT DAO formulation takes the following form:

$$
\begin{aligned}
\underset{\{f_{b\alpha}, c_{b\alpha}, u_{b\alpha}\}_{b,\alpha=1}^{n_b,n_\alpha}}{\textit{minimize}} \quad & \underbrace{\frac{1}{2}\left\| W\left(\left(\sum_{b=1}^{n_b}\sum_{\alpha=1}^{n_\alpha} A_{b\alpha} f_{b\alpha}\right) - d\right)\right\|_2^2}_{\textit{fidelity term}} \\
& + \sum_{b=1}^{n_b}\sum_{\alpha=1}^{n_\alpha} \underbrace{\left(\lambda_1 \left\|D_{1_{b\alpha}} f_{b\alpha}\right\|_1 + \lambda_2 \left\|D_{2_{b\alpha}} f_{b\alpha}\right\|_1\right)}_{\textit{anisotropic TV term on } f} \\
& + \frac{1}{2}\sum_{b=1}^{n_b}\sum_{\alpha=1}^{n_\alpha} \underbrace{\left(\gamma_1 \left\|\sqrt{diag(u_{b\alpha})}(f_{b\alpha} - c_{b\alpha})\right\|_2^2 + \gamma_2 \left\|\sqrt{diag(1 - u_{b\alpha})} f_{b\alpha}\right\|_2^2\right)}_{\textit{single segment term}} \\
& + \sum_{b=1}^{n_b}\sum_{\alpha=1}^{n_\alpha} \underbrace{\left(g_1 \left\|D_{1_{b\alpha}} u_{b\alpha}\right\|_1 + g_2 \left\|D_{2_{b\alpha}} u_{b\alpha}\right\|_1\right)}_{\textit{anisotropic TV term on } u} + \gamma_3 \sum_{b=1}^{n_b}\sum_{\alpha=1}^{n_\alpha} \underbrace{S_{b\alpha}(1 - P_{b\alpha})\|f_{b\alpha}\|_2}_{\textit{group sparsity term}} \\
& + \; g_3 \underbrace{\|D_P u\|_1}_{\textit{aperture continuity term}}
\end{aligned}
$$

$$
\textit{subject to} \qquad f_{b\alpha} \geq 0, c_{b\alpha} \geq 0, 0 \leq u_{b\alpha} \leq 1, b = 1,2,..,n_b, \alpha = 1,2,..,n_\alpha
$$

$$
u = \begin{bmatrix} u_{b=1\,\alpha=1}^T & u_{b=1\,\alpha=2}^T & \cdots & u_{b=1\,\alpha=n_\alpha}^T & u_{b=2\,\alpha=1}^T & u_{b=2\,\alpha=2}^T & \cdots & u_{b=n_b\,\alpha=n_\alpha}^T \end{bmatrix}^T
$$

Equation 4-9

where $f_{b\alpha}$, $c_{b\alpha}$ and $u_{b\alpha}$ are the optimization variables. Gantry angles are indexed by b, which ranges from 1 to n_b, collimator angles are indexed by α, which ranges from 1 to n_α. $f_{b\alpha}$ is the vectorized fluence map of the candidate beam at gantry angle b and collimator angle α. $c_{b\alpha}$ is a scalar quantity and only has one value per gantry angle and collimator angle, which is the fluence intensity that $f_{b\alpha}$ approaches within an aperture. $u_{b\alpha}$ is the aperture variable, defined as 1 where the aperture exists and 0 elsewhere.

The fluence to dose transformation matrix is denoted by $A_{b\alpha}$, and the desired dose d_0, is set as the prescription dose at the PTV and zero elsewhere. The weighting matrix W is a diagonal matrix, with weightings for structures of interest as diagonal elements. The derivative matrices, $D_{1_{b\alpha}}$ and $D_{2_{b\alpha}}$, take the derivative of the fluence map $f_{b\alpha}$ and aperture $u_{b\alpha}$ in both directions parallel and orthogonal to the MLC leaf movement. $P_{b\alpha}$ is a scalar that equals to 1 for selected on-beams (selected gantry and collimator angle) from collimator angle selection (CAS), and 0 elsewhere. $S_{b\alpha}$ weighs the group sparsity term for each candidate beam to avoid over-penalizing beams with long radiological pathlengths to the target by normalizing based on the $A_{b\alpha}$ matrix. The definition of $S_{b\alpha}$ is

$$S_{b\alpha} = \frac{sum\left(I_{PTV} A_{b\alpha} \vec{1}\right)}{\sqrt{n_{b\alpha}}}$$

where I_{PTV} is a diagonal indicator matrix for PTV, with its diagonal elements equal to 1 for voxels in PTV and 0 elsewhere, and $n_{b\alpha}$ is the number of beamlets with a trajectory that intersects PTV in beam at gantry angle b and collimator angle α. u is a concatenation of the aperture variables $u_{b\alpha}$ at all gantry angles and collimator angles. The derivative matrix D_P takes derivative on the apertures of adjacent selected on-beams.

In Equation **4-9**, the data fidelity term minimizes the differences between the calculated dose and the desired dose d. The anisotropic total variation (TV) regularization on fluence map $f_{b\alpha}$ encourages piecewise continuity of the fluence maps within each beam[130]. λ_1 and λ_2 are hyperparameters controlling the weightings of the TV regularization. A balance between dose fidelity and fluence map continuity improves final dose profile. The single segment term pushes $f_{b\alpha}$ towards $c_{b\alpha}$ where the aperture $u_{b\alpha}$ is 1, and towards 0 where $u_{b\alpha}$ is 0. It encourages fluence map to be a constant within the aperture of each beam, and 0 elsewhere. By weighing this term heavily, the optimizer forces only one aperture segment per beam. The anisotropic TV term on aperture variable $u_{b\alpha}$ encourages intact aperture shape and penalizes holes in the aperture. The group sparsity term is a $l_{2,1}$ norm penalty, which promotes group sparsity in the fluence map $f_{b\alpha}$ and inactivates most of candidate beams. Note that this term is controlled by $P_{b\alpha}$, which only enforces penalty on candidate beams that were not included as on-beams from CAS. The aperture continuity term regulates leaf movement by penalizing differences of apertures between adjacent selected on-beams.

4.4.2.2 Collimator angle selection

Collimator angle is selected by choosing the shortest path on the simplified Dijkstra's graph. The graph contains $n_\alpha * n_b$ nodes, where each node cost is associated with a candidate beam. The node cost $NC(b, \alpha)$ for the candidate beam with gantry angle b and collimator angle α depends on $f_{b\alpha}$, the fluence map from DAO at the current iteration. The formulation of $NC(b, \alpha)$ is given by

$$NC(b, \alpha) = \frac{1}{2} \|W(A_{b\alpha} f_{b\alpha} - d)\|_{2.}^{2}$$

Equation 4-10

Candidate beams that correspond to dose contribution closer to the desired dose distribution will be assigned with lower node costs. By finding the shortest path on the Dijkstra's graph, optimal candidate beams are selected.

The edge costs between every two nodes enforce constraints on trajectory selection. The edge cost between (b_1, α_1) and (b_2, α_2) is defined as

$$EC\big((b_1, \alpha_1), (b_2, \alpha_2)\big) = \begin{cases} 0 & if\ b_2 - b_1 = 1\ and\ min(\|\alpha_1 - \alpha_2\|, 180^\circ - \|\alpha_1 - \alpha_2\|) < \alpha_0 \\ \infty & otherwise \end{cases}$$

Equation 4-11

Equation 4-11 ensures that the collimator angles of adjacent selected on-beams are within collimator rotation speed limits (controlled by α_0), and that only one collimator angle is selected per gantry angle. The selected on-beams are denoted by α_{short}, a vector with n_b

elements, and each element $\alpha_{short}(b)$ represents the selected collimator angle for gantry angle b, ranging from the first gantry angle ($b = 1$) to the last gantry angle ($b = n_b$).

Even after regularization using the group sparsity term, there can be multiple collimator angles co-existing in some gantry angles. To allow multiple collimator angles, a few 'slow beams' are selected apart from the shortest path α_{short} on the Dijkstra's Map. These slow beams $\{b_1, b_2, \cdots, b_N\}$ are chosen as the gantry angles that correspond to the N lowest sum-node-costs, which is defined as

$$SNC(b) = \sum_{\alpha=1}^{n_\alpha} NC(b,\alpha) - NC\big(b, \alpha_{short}(b)\big)$$

When delivering these 'slow beams', the gantry slows down to allow a full cycle of collimator rotation.

4.4.2.3 Algorithm

4.4.2.3.1 Direct Aperture Optimization using FISTA

The algorithm that we use to solve the DAO in Equation **4-9** cycles among 3 modules, each of which minimizes with respect to one block of variables (either f, c, or u) while holding the other variables constant. Within each module, the optimization problem is solved using FISTA [27,155]. The Huber penalty[156] was used as a smooth approximation of the l_1 norm to facilitate the application of FISTA.

4.4.2.3.2 Collimator Angle Selection: Dijkstra's Algorithm

Collimator angles are selected by choosing the shortest path on a simplified Dijkstra's graph with n_α rows and n_b columns, including $n_\alpha * n_b$ nodes with node costs and edge costs defined in equation (2) and equation (3). Equation (3) ensures that the shortest path is chosen so that it starts from the first column ($b = 1$) and ends at the last column ($b = n_b$), and that only one node is selected per column. Algorithm 2 presents the graph optimization algorithm to find the shortest path α_{short} on the simplified Dijkstra's graph with such layered structure of rows and columns. α_{short} will be a vector with n_b elements denoting the selected collimator angle for each gantry angle.

Once α_{short} is chosen, the sum-node-costs are evaluated based on equation (4), and slow beams are selected. $P_{b\alpha}$ is assigned to be 1 for any selected candidate beams and 0 elsewhere. This variable will be utilized in DAO to discourage the development of any unselected beams.

Simplified Dijkstra's algorithm

```
Initialize the graph with each node $G(b,\alpha) \coloneqq \infty$
$G(1,\alpha) \coloneqq NC(1,\alpha)$
for $b = 1, 2, \dots, n_b - 1$
    for $\alpha = 1, 2, \cdots, n_\alpha$
        $G(b+1,\alpha) \coloneqq \min_{\alpha'} G(b,\alpha') + NC(b+1,\alpha) + EC\big((b,\alpha'),(b+1,\alpha)\big)$
        $\alpha_t(b+1,\alpha) \coloneqq \operatorname{argmin}_{\alpha'} G(b,\alpha') + NC(b+1,\alpha) + EC\big((b,\alpha'),(b+1,\alpha)\big)$
    end for
end for
$\alpha_{short}(n_b) \coloneqq \operatorname{argmin}_{\alpha'} G(n_b,\alpha')$
for $b = n_b - 1, n_b - 2, \dots, 1$
    $\alpha_{short}(b) \coloneqq \alpha_t\big(b+1, \alpha_{short}(b+1)\big)$
end for
```

Table 4-9 Pseudocode for Simplified Dijkstra's algorithm

4.4.2.4 Evaluation

The algorithm was first tested on a spherical phantom with a diameter of 10 cm, as shown in Figure 4-22. Two spherical PTVs, with a diameter of 1.3 cm and 1.7 cm, respectively, are distributed symmetrically on the axial plane with a distance of 2.25 cm. The phantom body is the only OAR in this case. The simple geometric configuration provides a controlled test environment for the proposed optimization framework, which focuses on the effects of collimator rotation and rules out other factors that may contribute to the dose outcome, such as the tradeoff among multiple OARs. This phantom is also clinically relevant to multiple brain metastases. The algorithm was also assessed on a glioblastoma multiforme patient (GBM), a lung cancer patient (LNG), and a prostate cancer patient (PRT). Table 4-10 shows the respective prescription doses and PTV volumes in all cases.

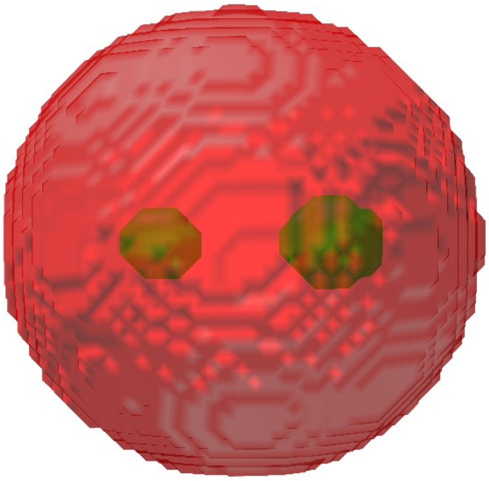


Figure 4-22 3D plot of the phantom.

	Prescription Dose (Gy)	**PTV Volume (cc)**
phantom	50	3.77
GBM	25	6.23
LNG	50	138.50
PRT	40	111.16

Table 4-10 Prescription doses and PTV volumes

The dose calculation uses a convolution/superposition code with a 6 MV x-ray polyenergetic kernel, described in our previous publications[100,107], calculated for 60 gantry angles with 6 degrees of separation and 10 collimator angles with 18 degrees of separation. The ratio of gantry and collimator angular resolution was determined based on the relative rotational speed of the collimator and gantry ($\frac{\omega_{collimator}}{\omega_{gantry}} = 2.5$). This ratio can be adjusted based on the actual machine specification. The dose array resolution was 0.25×0.25×0.25 cm^3. For the phantom study, the SC-VMAT plans were generated for both 0.5×0.5 cm^2 and 0.25×0.25 cm^2 beamlet resolution and were compared with the DC-VMAT plan with 0.5×0.5 cm^2 beamlet resolution. In patient cases, the beamlet resolution was 1×1 cm^2 due to the limited size of the computation.

In the phantom study, both the DC-VMAT and the SC-VMAT plans include one coplanar arc. In patient studies, the DC-VMAT plan results in one coplanar arc with 13 slow beams where the gantry slows down to allow a full collimator rotation, and is compared against the SC-VMAT plan with three coplanar arcs with collimator angles at 60º, 120º, 180º, respectively. The SC-VMAT plan utilizes the same optimization model except that collimator angles are fixed within each arc. The traveling time between each beam is 1.2s, allowing 18º of collimator rotation.

Evaluation includes PTV D95, D98, D99, Dmax, PTV homogeneity, defined as $\frac{\mathrm{D95}}{\mathrm{D5}}$, and the dose conformity, defined as the overlap of the 100% isodose volume and the PTV. For the phantom study, R10-R90 was evaluated to quantify the dose compactness. For the patient studies, the organs-at-risk (OAR) Dmax and Dmean were assessed. Max dose is defined as the dose at 2% of the structure volume, D2, recommended by the ICRU-83 report.

4.4.2.5 Implementation Details

To achieve a good balance between dosimetry quality and plan deliverability, there are 8 additional parameters in equation (1) that need to be tuned based on the following considerations. λ_1 and λ_2 were tuned to smooth the fluence map while keeping its major structures. The same principle applied to g_1, g_2, and g_3, controlling the smoothness of the aperture variable $u_{b\alpha}$. γ_1 and γ_2 are controlled by the equations

$$\gamma_1 = \gamma_{10}\left(1 + 1000 * \left(\frac{n}{N}\right)^8\right)$$

$$\gamma_2 = \gamma_{20}\left(1 + 1000 * \left(\frac{n}{N}\right)^8\right),$$

where n is the current iteration number, and N is the total number of iterations. The monotonically increasing fashion of γ_1 and γ_2 allows the optimization to focus on the convex dose fidelity term at the early stages of the optimization and then enforces deliverability during the late iterations. The parameter γ_3 is automatically adjusted in optimization to achieve the desired number of on-beam.

The parameters used in this study are summarized in Table 4-11. The table gives the initial values of γ_1 and γ_2 which monotonically increase, and γ_3 which is automatically adjusted to achieve the desired number of on-beams. Although the parameter tuning appears to be tedious and empirical, based on our experience, the optimization results are relatively insensitive to these parameters within two orders of magnitude. This is particularly true for patients of the same disease sites where the same set of optimization parameters appear to be adequate without the need for individual tuning.

	λ_1	λ_2	γ_{10}	γ_{20}	γ_3	g_1	g_2	g_3
phantom	30000	10000	5000	5000	2000	0.008	0.005	0.008
GBM	10000	5000	1000	1000	2000	0.05	0.01	0.01
LNG	500000	100000	50	50	5000	0.5	0.1	0.1
PRT	100	20	5000	5000	1000	0.05	0.01	0.01

Table 4-11 Parameters setup for all cases

To accelerate the algorithm, the system matrix A is downsampled in both the fluence map domain and the patient voxel domain. The downsampling in the fluence map domain is performed by removing candidate beams that are inactive throughout the DAO module. In the patient voxel domain, both PTV and OARs are downsampled until collimator angles converge. To avoid geometry deformation of these critical structures through downsampling, we keep a high sampling rate on the outer layer of each structure and a low sampling rate on the interior of each structure. After the collimator angles converge, the DAO is performed with fixed collimator angles and a full sampling of PTV and OARs.

4.4.3 Results

4.4.3.1 Phantom study

With the same target coverage, the DC-VMAT is able to remarkably enhance the dose compactness. Table 3 shows the comparable PTV statistics of the DC-VMAT plan using 5mm MLC leaf thickness and the SC-VMAT plan using 5mm/2.5mm MLC leaf thickness. Figure 4-23 shows the percentage reduction in the R10-R90 of the DC-VMAT (5mm) plan and the SC-VMAT (2.5mm) plan compared with the SC-VMAT (5mm) plan. The reduction in dose spillage is remarkably similar between the DC-VMAT (5mm) plan and the SC-VMAT (2.5mm) plan, suggesting that the DC-VMAT plan achieves an effectively higher modulation resolution than its physical leaf thickness. Specifically, DC-VMAT (5mm) plan reduced SC-VMAT (5mm) plan R50 by 20.3% with the same MLC leaf thickness. The improvement in dose compactness using DC-VMAT (5mm) and SC-VMAT (2.5mm) in comparison to SC-VMAT (5mm) can also be visualized in the isodose line plots (Figure 4-24) where the 24Gy, 32Gy, and 40Gy isodose lines more tightly hug the targets.

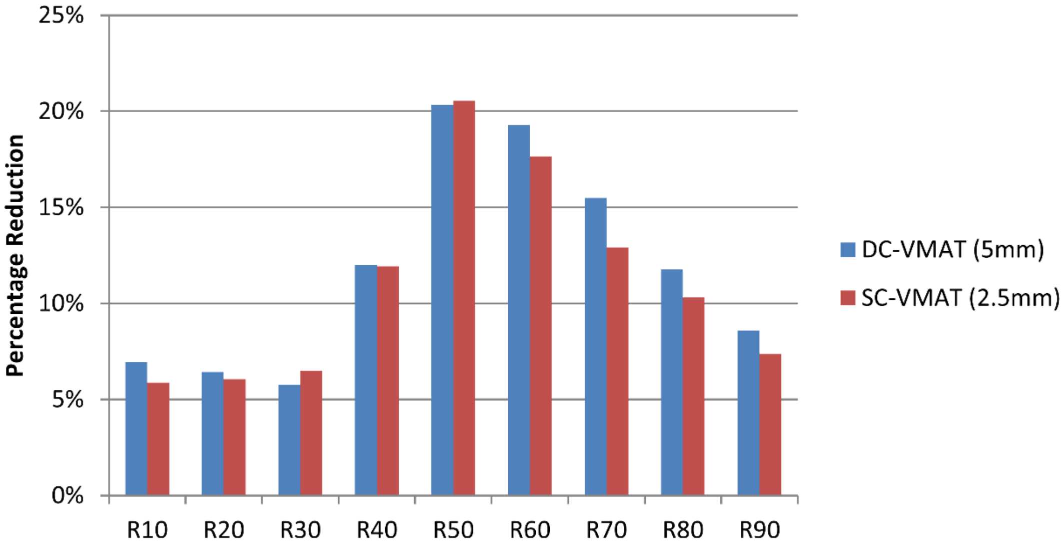


Figure 4-23 Percentage reduction in R10-R90 of the DC-VMAT (5mm) plan and the SC-VMAT (2.5mm) plan, compared with the SC-VMAT (5mm) plan on the phantom study.

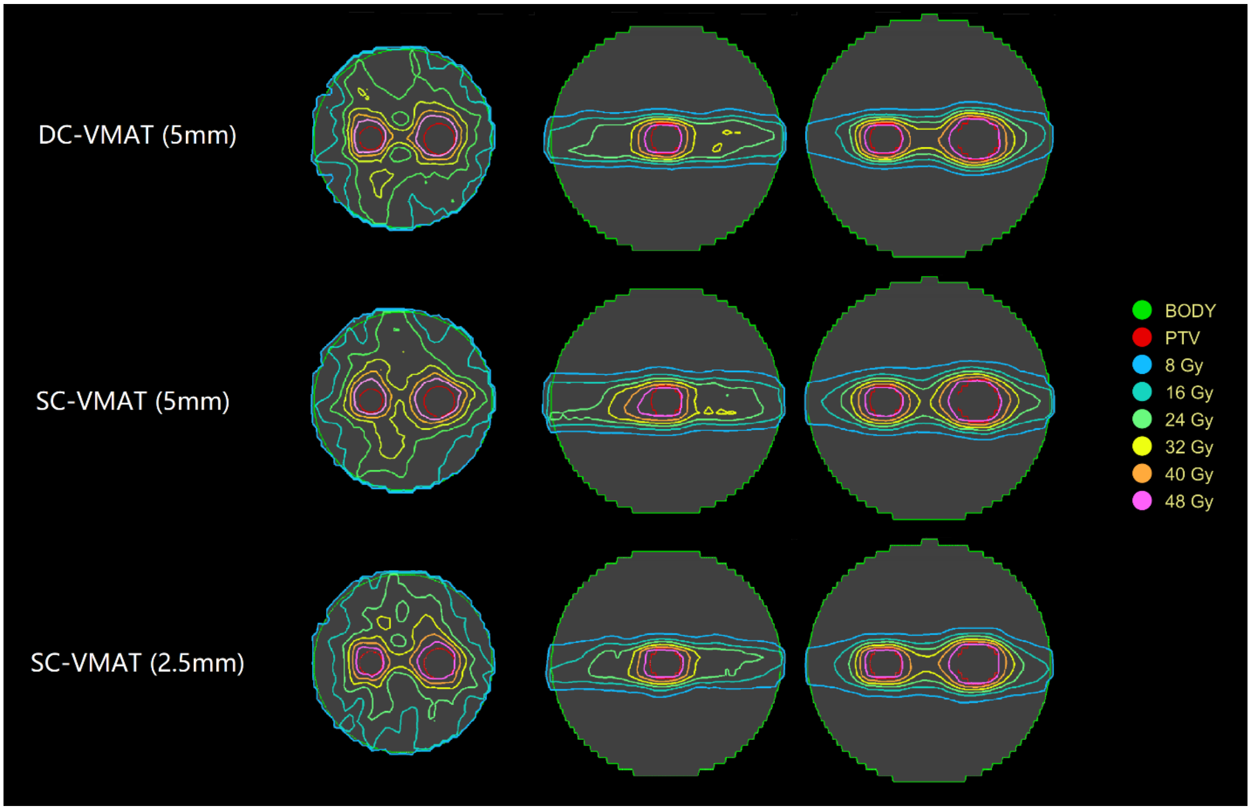


Figure 4-24 Isodose plots for the phantom.

	DC-VMAT (5mm)	SC-VMAT (5mm)	SC-VMAT (2.5mm)
Dmax	52.62	52.58	52.66
D95	50.00	49.99	50.01
Dmin	49.30	49.36	49.37
D99	48.75	48.90	49.05
Dmean	51.64	51.46	51.49
Homogeneity	0.95	0.95	0.95
dose conformity	0.95	0.95	0.95

Table 4-12 PTV statistics for the phantom

4.4.3.2 Patient studies

Figure 4-25 is a scheme of the fluence map for DC-VMAT plan. Each row corresponds to a collimator angle, at a separation of 18º, and each column represents a gantry angle, at 6º of separation. The DC-VMAT method optimizes all candidate beams simultaneously and selects the optimal collimator angles for each gantry angle. 13 beams are chosen as slow beams, indicated by the columns with all beams on. The maximum allowance of the collimator rotation is 18º between two adjacent beams. With the collimator rotation speed limit at 15º per second, the plan takes around 3.5 minutes in total to deliver.

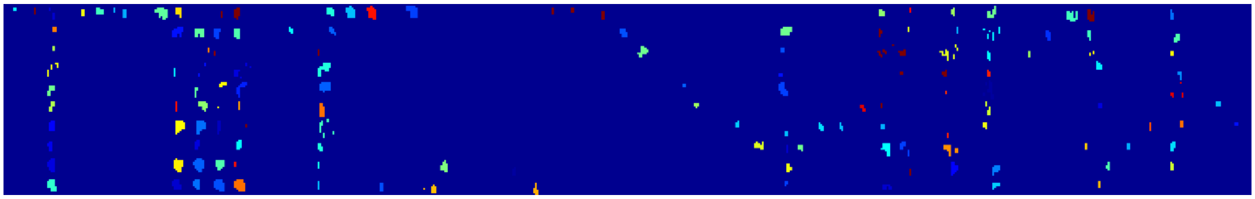


Figure 4-25 Scheme of the fluence map for all candidate beams for LNG case. MLC leaf direction is vertical for this diagram. The columns with all beams on represent the 'slow beams'.

Figure 4-26 shows the dose distribution of DC-VMAT and SC-VMAT for all patients. By utilizing the new optimization freedom, DC-VMAT was able to achieve better dose sparing on OARs while keeping the same or better PTV coverage. In comparison to SC-VMAT, for the GBM patient, the DC-VMAT was able to reduce dose spillage to the brainstem, chiasm, and right eye, whereas the SC-VMAT covered these OARs with up to 3 Gy of dose. In the LNG case, the DC-VMAT substantially reduced the dose to the ribs near PTV, spinal cord, esophagus, and trachea. In the prostate case, DC-VMAT reduced dose to the rectum, R/L Femur, and penile bulb.

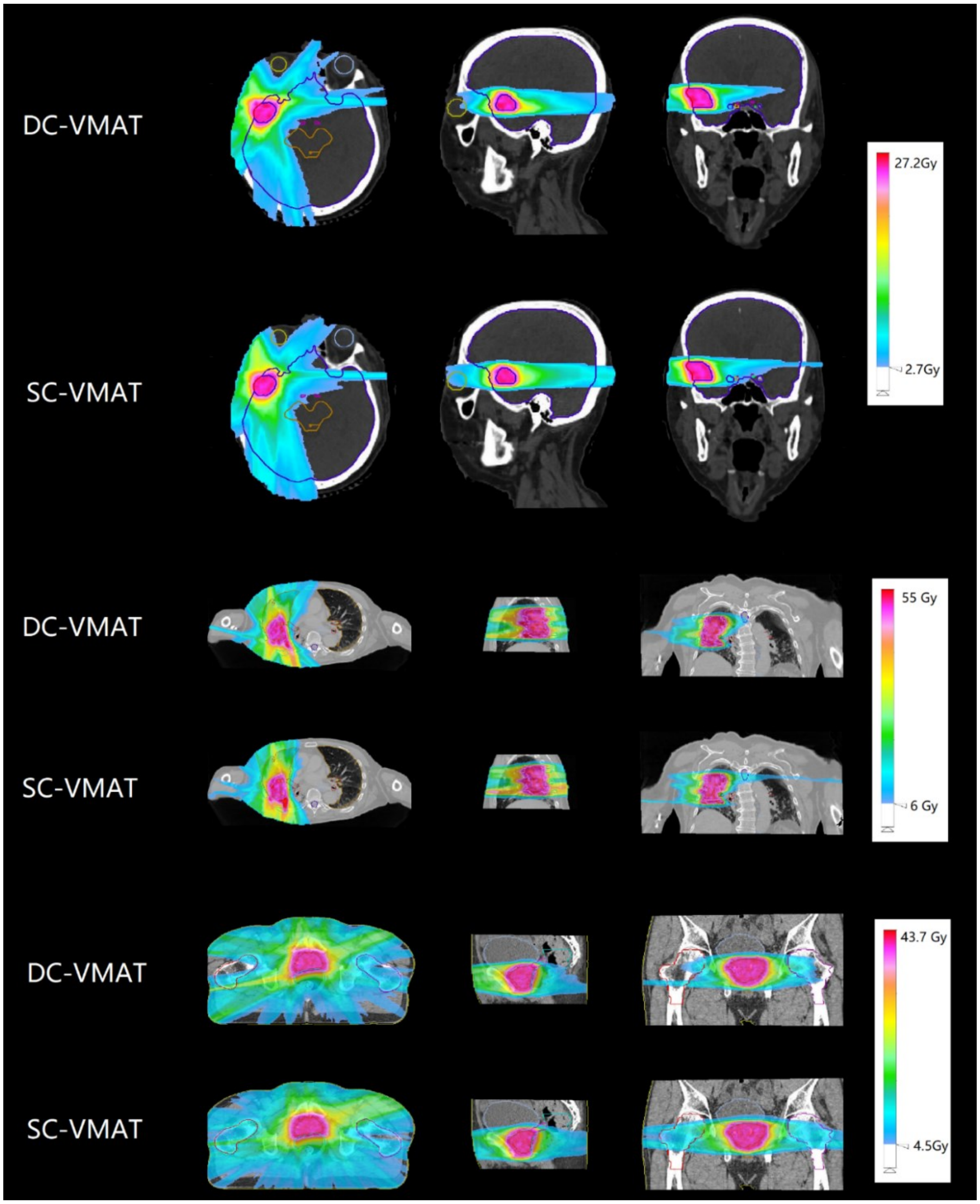


Figure 4-26 Dose washes for all patients. The low dose cutoff is at around 10% of the prescription dose.

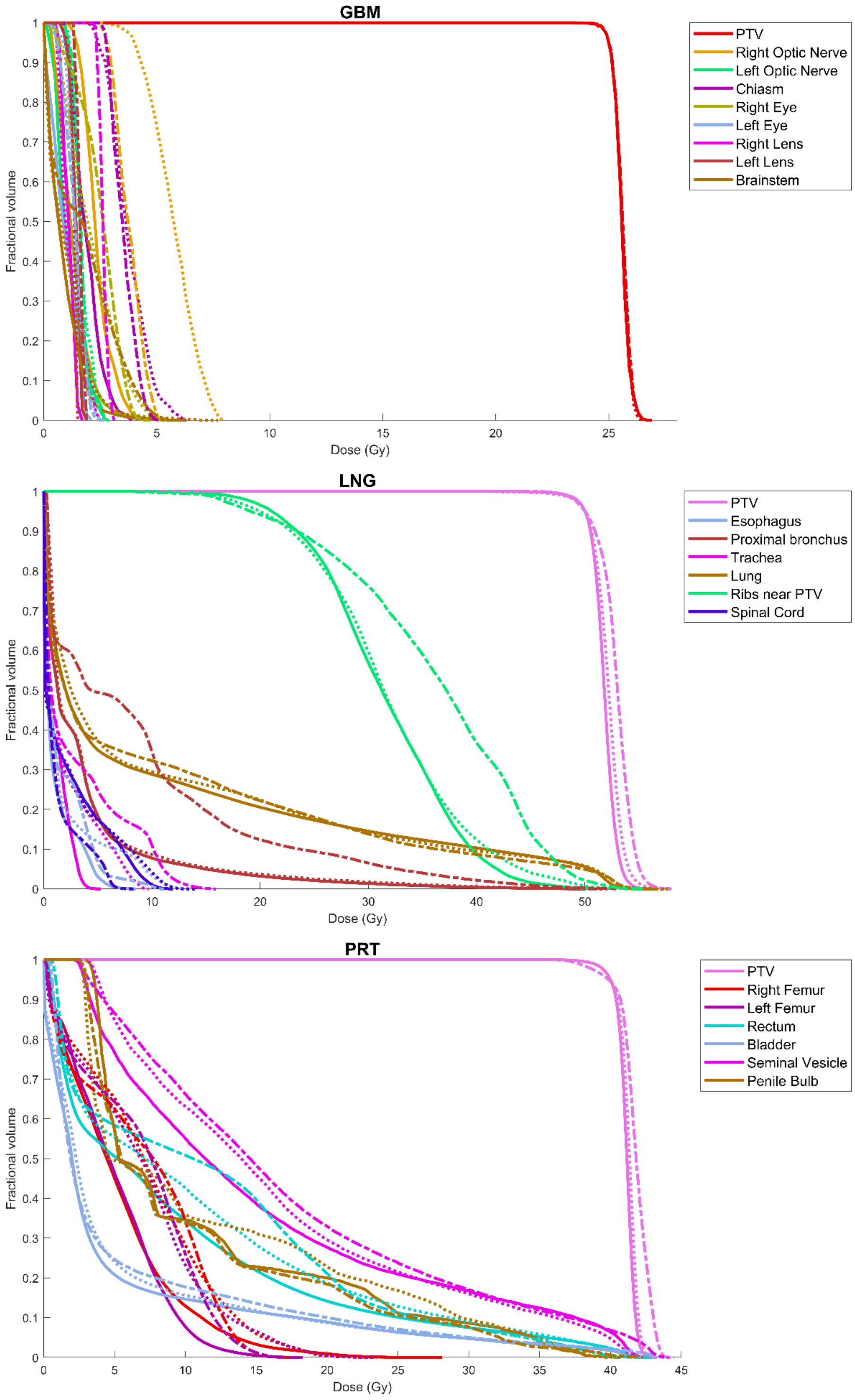


Figure 4-27 DVH comparison of the DC-VMAT with 10 mm MLC leaf thickness (solid line), SC-VMAT with 10 mm MLC leaf thickness (dotted line), and clinical VMAT plan with 5 mm MLC leaf thickness (dashed line) for all patients. D95 is normalized to the prescription dose.

Figure 4-27 shows the DVHs of DC-VMAT, SC-VMAT, and clinical VMAT for all patients. The clinical VMAT plan using 5 mm MLC leaf thickness was included as a comparison with the DC-VMAT and SC-VMAT plans using 10 mm MLC leaf thickness. In the GBM case, where the PTV size is comparable to the collimator resolution, the clinical VMAT shows improved dose profile as compared with SC-VMAT, due to the benefits of using fine resolution MLC. The DC-VMAT, on the contrary, was able to overcome the limitation of using coarse resolution MLC and remarkably improve the dose profile as compared with SC-VMAT, achieving an even more conformal dose than the clinical VMAT plan with fine resolution MLC. In the LNG and PRT case, where the effects of collimator resolution are limited, both the DC-VMAT and SC-VMAT substantially improved the dose profile compared with the clinical VMAT plan, in agreement with our previous study[7]. The PTV statistics of the DC-VMAT plan and SC-VMAT plan were comparable, while the DC-VMAT was able to globally reduce the OARs dose.

The quantitative statistics for PTV and OARs of the DC-VMAT and SC-VMAT plans are shown in Table 4 and Table 5, respectively. On average, the DC-VMAT plan reduced the OARs max and mean dose by 4.49% and 2.53% of the prescription dose, with the same target coverage. The single largest sparing in OARs max and mean dose is up to 15% and 13.5% of the prescription dose.

Patient Case	PTV Statistics							
	Homogeneity		Dose conformity		D95	D98	D99	Dmax
	DC-VMAT	SC-VMAT	DC-VMAT	SC-VMAT	DC-VMAT – SC-VMAT (Gy)			
GBM	0.956	0.958	0.949	0.947	-0.009	+0.018	+0.078	+0.085
LNG	0.936	0.925	0.95	0.95	-0.004	+0.183	+0.536	-0.789
PRT	0.953	0.949	0.95	0.95	-0.009	+0.017	-0.032	-0.170

Table 4-13 PTV statistics for all patients

DC-VMAT – SC-VMAT (Gy)	Dmax		Dmean	
	Largest Value	Average Value	Largest Value	Average Value
GBM	−3.76 R Opt Nrv	−1.12	−3.38 R Opt Nrv	−0.90
LNG	−5.28 Esophagus	−3.08	−0.83 Trachea	−0.45
PRT	-5.44 L Femur	-1.13	-1.91 R Femur	-1.23

Table 4-14 OARs statistics for all patients

4.4.4 Discussion

The non-progressive-sampling comVMAT model[35] in our previous study optimizes all beams simultaneously instead of progressively inserting new beams, and solves the VMAT optimization problem without relying on greedy heuristics, achieving both theoretical and practical advantages. The DC-VMAT model extends the comVMAT model, and incorporates dynamic collimator rotation during the arc. A novel group sparsity term to optimize the collimator/gantry angles and Dijkstra's term to select the optimal collimator angle for each individual gantry angle were used to incorporate this new freedom. An alternating strategy between the DAO module and the collimator angle selection module was adopted to achieve the final selection of optimal beams.

As a result, DC-VMAT is able to achieve a higher effective collimator resolution and subsequently improves dosimetry compared with SC-VMAT, which was shown superior to the existing progressive sampling VMAT method[35]. The superior fluence modulation resolution is particularly demonstrated by the phantom study. Under the controlled condition without tradeoffs among multiple OARs, the improvement in dose compactness using a dynamic collimator with standard resolution MLC matched that of the static collimator with high-resolution MLC. The improved effective MLC resolution may offer a compelling solution to the difficult decision a radiation oncology department often has to make between MLC leaf resolution and field-of-view.

This current study is different from previous collimator rotation studies[152,153] in both theory, implementation and results. The previous methods align the collimator angle with organs of distinct geometrical orientations, such as the spinal cord. Because this operation is not included in the optimization, the optimality of such an approach cannot be proved. Furthermore, in most clinical cases, such as the ones presented in this paper, there is no such clear-cut geometry for the planner to set up collimator angles in the beams-eye-view. Therefore, we believe our method incorporating the collimator rotation term in optimization is more broadly applicable.

An advantage of collimator rotation is that compared to other additional optimization freedoms such as the non-coplanar angles that have been recently researched, collimator rotation is straightforward to implement as it does not increase the risk of collision. For the same reason, there is a potential to safely accelerate collimator rotation for more rapid DC-VMAT delivery. The current collimator rotation speed is limited at 15º/s, and subsequently, the estimated delivery time is 72s for the DC-VMAT one-arc plan and 60s for the SC-VMAT one-arc plan in the phantom study, and around 200s for the DC-VMAT one-arc plan with slow beams and 180s for the three-arc SC-VMAT plan in the patient studies. The delivery time for DC-VMAT could be further reduced with accelerated collimator rotation.

In the patient studies, both the gantry angle and MLC resolution is lower than typical clinical parameters due to two implementation considerations. First, the group sparsity terms on the collimator rotation will penalize similar collimator angles for adjacent beams because they contribute similarly to the dosimetry. This resulted in large collimator rotation between beams. If the gantry angles are more densely sampled, the entire treatment has to slow down to wait for the collimator rotation. The relatively coarse gantry angle gives enough time for the collimator to rotate without having to slow down at most gantry angles. This constraint would be relaxed with the aforementioned faster collimator rotation. The second reason is that DC-VMAT is computationally intensive due largely to the size of the optimization problem involving multiple collimator rotations for a gantry angle. The vast problem size is compounded by the alternating optimization steps between DAO and CAS. To mitigate the computational challenge, the DAO component is now solved using FISTA, which is computationally inexpensive and effectively solves the VMAT optimization as demonstrated in Section 4.2. However, even with the accelerated algorithm, the DC-VMAT optimization run time varied from 30 minutes for the GBM case to 3 hours for the LNG case and PRT case, whereas the SC-VMAT optimization run time ranged from 5 minutes for the GBM case to 30 minutes for the LNG and PRT case. To further improve the computational performance and use higher MLC resolution, other than moving the calculation from MATLAB to a higher performing language, the optimization may be performed on a parallel computing architecture, where the most computationally expensive matrix multiplication step in FISTA can be parallelized.

Another limitation of this study is that DVH constraints are not used in optimization to maintain convexity in the fidelity term. While DVH constraints can be convenient in clinical practice, introducing them will make the optimization problem more non-convex, less stable, and slower to solve. The current optimization framework was compared with DVH based planning method in Figure 4-27 and in our previous study[7] and showed consistently better dosimetric results, proving that the voxel-based optimization method is not a major hindrance to achieve high quality plans. The main point of this study is separate from the DVH constraint consideration. We would, of course, investigate the feasibility and implications of adding such terms in future studies.

4.4.5 Summary

The dynamic collimator rotation effectively increases the modulation resolution and improved dose compactness and OAR sparing compared with static collimator VMAT (SC-VMAT).

4.5 A novel optimization framework for VMAT with dynamic gantry couch rotation

4.5.1 Introduction

Current VMAT methods utilize one or more arcs, each co-planar by itself, with or without manually selected couch rotations *between* arcs. For clarity, we term it 2πVMAT in this paper. The 2πVMAT optimization problem has been solved by progressively inserting new control points into coarsely sampled gantry angles, until typically a 2-degree spacing between the gantry angles is achieved. This progressive sampling approach is not only computationally tractable, but also provides deliverable VMAT plans that meet the mechanical constraints of MLC, including the leaf travelling speed limits. However, this heuristic progressive sampling approach has several limitations. First, the greedy optimization algorithm is susceptible to undesired local minima due to its sensitivity to suboptimal parameter tuning in the early coarse-resolution stage of optimization. Second, the optimization result strongly depends on the parameter tuning history, causing difficulties to exactly reproduce a plan. Third, due to the interpolation that take place in the progressive sampling process, the method tends to spread the photon fluence uniformly throughout the entire arc even if only a fraction of the beam angles is optimal for treating the patient. To overcome these limitations, we developed a level-set based DAO for 2πVMAT, which solves the entire arc optimization problem in full resolution. This non-progressive sampling approach was shown to generate a single arc coplanar VMAT that outperformed progressive sampling VMAT using two arcs with the same number of control points in each arc.

The improved VMAT method is still limited to coplanar arcs while non-coplanar beams hold unquestionable dosimetric advantages as shown by recent 4π IMRT research, which maximally utilizes the non-coplanar beams for significant dosimetric gains compared to the VMAT plans. The advantage of non-coplanar IMRT beams is further demonstrated by Sharfo *et al.* [157] who combined VMAT with a few non-coplanar IMRT beams and showed improved organs-at-risk (OAR) sparing and dose spillage. However, delivering many isolated non-coplanar beams can be time-consuming and laborious to the clinic and an undue burden to the patient. Alternatively, non-coplanar VMAT arcs were investigated.

A straightforward non-coplanar VMAT approach was based on user-defined trajectories [118,148,158,159] that are limited to narrow applications or specific patient anatomies. Several automated VMAT trajectory optimization techniques were proposed to generalize the solution. Smyth *et al.* [113] introduced a trajectory optimization for VMAT by choosing a minimal cost trajectory on a cost map, which was computed based on ray-OAR voxel intersections. Yang *et al.* [160] proposed a hierarchical clustering algorithm to find multiple continuous and extended sub-arcs through a minimum search of a score function containing geometrical information. MacDonald *et al.* [161] also introduced a trajectory optimization algorithm that minimizes the geometric overlap between planning target volumes (PTV) and OARs based on the two-dimensional projection from source to the isocenter plane as a function of gantry and couch angle. VMAT optimization was then performed as a separate step. Due to the separation, these methods were ineffective in solving complex planning problems where clean trajectories not entering one or more OARs do not exist. To overcome the limitation, a gantry/couch trajectory optimization needs to be incorporated in dose optimization.

Due to the difficulty of solving the complete non-coplanar VMAT problem, a mathematically tractable way is to first identify the non-coplanar control points using beam orientation optimization (BOO) and fluence map optimization (FMO), which essentially is the goal of 4π IMRT. Following the idea, Wild *et al.*[162] utilized a genetic algorithm to solve the combinatorial problem of BOO, and Papp *et al.*[163] used the gradient norm strategy to heuristically select a few promising beams. Once the static beam positions are determined as nodes, non-coplanar arcs are created to connect them. An intrinsic limitation of these methods is that although the static beam positions are dosimetrically desirable, the arcs connecting them are not. By generating the non-coplanar arc plans, the major workload of dose delivery is shifted to these dosimetrically suboptimal arc trajectories.

Evidently, to fundamentally solve the non-coplanar VMAT problem, not only the nodes, but also entire arc trajectories need to be part of the BOO equation. In this study, we propose a novel optimization framework that simultaneously solves the complete non-coplanar VMAT trajectory optimization and DAO problems for VMAT, while ensuring deliverability by avoiding couch-gantry-patient collision and enforcing mechanical constraints of MLC leaf motion and gantry rotation. We term this method 4πVMAT in contrast to the 2πVMAT methods using only individual coplanar arcs.

4.5.2 Methods

4.5.2.1 Direct Aperture Optimization and Beam Orientation Optimization

The proposed 4πVMAT DAO and BOO formulation is written as

$$
\begin{aligned}
\underset{\{f_{b\alpha}, c_{b\alpha}, u_{b\alpha}\}_{b,\alpha=1}^{n_b,n_\alpha}}{\textit{minimize}} \quad & \underbrace{\frac{1}{2}\left\| W\left(\left(\sum_{b=1}^{n_b}\sum_{\alpha=1}^{n_\alpha} A_{b\alpha} f_{b\alpha}\right) - d_0\right)\right\|_2^2}_{\textit{fidelity term}} \\
& + \sum_{b=1}^{n_b}\sum_{\alpha=1}^{n_\alpha} \underbrace{\left(\lambda_1 \left\| D_{1_{b\alpha}} f_{b\alpha} \right\|_1 + \lambda_2 \left\| D_{2_{b\alpha}} f_{b\alpha} \right\|_1\right)}_{\textit{anisotropic TV term on f}} \\
& + \frac{1}{2}\sum_{b=1}^{n_b}\sum_{\alpha=1}^{n_\alpha} \underbrace{\left(\gamma_1 \left\| \sqrt{diag(u_{b\alpha})}(f_{b\alpha} - c_{b\alpha}) \right\|_2^2 + \gamma_2 \left\| \sqrt{diag(1 - u_{b\alpha})} f_{b\alpha} \right\|_2^2\right)}_{\textit{single segment term}} \\
& + \sum_{b=1}^{n_b}\sum_{\alpha=1}^{n_\alpha} \underbrace{\left(g_1 \left\| D_{1_{b\alpha}} u_{b\alpha} \right\|_1 + g_2 \left\| D_{2_{b\alpha}} u_{b\alpha} \right\|_1\right)}_{\textit{anisotropic TV term on u}} \\
& + \sum_{b=1}^{n_b}\sum_{\alpha=1}^{n_\alpha} \underbrace{\gamma_3 G_{b\alpha} \| f_{b\alpha} \|_2 + \gamma_4 G_{b\alpha}(1 - P_{b\alpha}) \| f_{b\alpha} \|_2}_{\textit{group sparsity term}} + g_3 \underbrace{\| D_P u \|_1}_{\textit{aperture continuity term}} \\
\textit{subject to} \quad & f_{b\alpha} \geq 0, c_{b\alpha} \geq 0, 0 \leq u_{b\alpha} \leq 1, b = 1, 2, .., n_b, \alpha = 1, 2, .., n_\alpha \\
& f_{b\alpha} = 0, c_{b\alpha} = 0, u_{b\alpha} = 0, \quad \forall (b, \alpha) \notin S \\
u = \big[u_{b=1\,\alpha=1}^T \quad & u_{b=1\,\alpha=2}^T \quad \cdots \quad u_{b=1\,\alpha=n_\alpha}^T \quad u_{b=2\,\alpha=1}^T \quad u_{b=2\,\alpha=2}^T \quad \cdots \quad u_{b=n_b\,\alpha=n_\alpha}^T\big]^T,
\end{aligned}
$$

Equation 4-12

where the notations for the variables and data are summarized in Table 4-15.

In this formulation, the data fidelity term attempts to find the optimal fluence map $f_{b\alpha}$ such that the total calculated dose from all candidate beams is as close as possible to the optimal dose d_0. The priorities for structures of interest are controlled by the diagonal weighting matrix W. The anisotropic total variation (TV) regularization on fluence map $f_{b\alpha}$ encourages piecewise continuity of the fluence map within each candidate beam [130]. λ_1 and λ_2 control the degree of piecewise continuity of $f_{b\alpha}$ in the direction parallel and orthogonal to the MLC leaf direction, respectively. A balanced tradeoff between dose fidelity and fluence map continuity achieves a high quality and deliverable dose distribution.

The single segment term enforces the final fluence map $f_{b\alpha}$ to contain only one segment per candidate beam. It pushes $f_{b\alpha}$ towards $c_{b\alpha}$ where $u_{b\alpha}$ is 1, and towards 0 where $u_{b\alpha}$ is 0, encouraging the fluence map to be a constant within the aperture, and 0 outside the aperture. By gradually increasing the weighting on this term during the optimization process, the optimizer forces single segment per candidate beam. The anisotropic TV term on aperture variable $u_{b\alpha}$ encourages large segments and penalizes holes in the aperture, with hyperparameters λ_1 and λ_2 controlling the degree of segment continuity in the direction parallel and orthogonal to the MLC leaf direction respectively. The single segment term and the anisotropic TV term address the hardware constraints by enforcing a single deliverable segment within each candidate beam.

The group sparsity term is a $l_{2,1}$ norm penalty. This convex penalty provides a non-greedy approach for BOO in 4πVMAT by promoting group sparsity in the fluence map $f_{b\alpha}$ and encouraging most of candidate beams to be inactive. This term is divided into two components through a parameter $P_{b\alpha}$, which is defined as 1 for candidate beams that are on the selected trajectory in the beam trajectory selection (BTS) process, and 0 elsewhere. γ_3 controls the sparsity level for all candidate beams and γ_4 adds additional weighting on candidate beams that are not on the optimal trajectory. The aperture continuity term regulates leaf movement by penalizing apertures differences between adjacent candidate beams on the selected trajectory from BTS. The group sparsity term and the aperture continuity term address the hardware constraints by switching off the off-trajectory beams by the end of the optimization and encouraging aperture continuity along the trajectory respectively.

The flowchart of the proposed 4πVMAT optimization framework is shown in Figure 4-28. First of all, a patient-specific collision map is generated and candidate beams are divided into a feasible set S and an infeasible set based on an individualized collision model for non-coplanar radiotherapy delivery [133]. Then $P_{b\alpha}$ is assigned as 1 for all feasible beams, which serves as prior information for the DAO&BOO. The DAO&BOO solves the optimization problem in Equation 4-12 and optimizes fluence map $f_{b\alpha}$, fluence intensity $c_{b\alpha}$, and vectorized aperture variable $u_{b\alpha}$ alternatingly, generating a 4πVMAT plan with a small fraction of beams active. Note that these on-beams are not necessarily on a connected trajectory; instead, they are separated in the entire feasible space. The optimal fluence map $f_{b\alpha}$ is then utilized to generate the simplified Dijkstra's map [154,164], on which the BTS is performed and one or two trajectories are selected based on the tumor region. $P_{b\alpha}$ is assigned as 1 for all candidate beams on the selected trajectories and 0 elsewhere, which is then translated into the DAO&BOO in the next iteration, penalizing heavily on the fluence map development for off-trajectory candidate beams as compared with on-trajectory beams. The optimization alternates between DAO&BOO and BTS until convergence, allowing BTS to fully explore the dose domain before converging to a final trajectory.

Notation	**Type**	**Description**
Indices		
b	Index	Index for couch angle, $\mathrm{b} = 1,2,..,\mathrm{n_b}$
α	Index	Index for gantry angle, $\alpha = 1,2,..,n_\alpha$
Optimization Variables		
$f_{b\alpha}$	Vector	Vectorized fluence map of the (b_{th}, α_{th}) candidate beam at couch angle b and gantry angle α
$c_{b\alpha}$	Vector	Fluence intensity that $f_{b\alpha}$ approaches within the aperture of the (b_{th}, α_{th}) candidate beam
$u_{b\alpha}$	Vector	Aperture variable for the (b_{th}, α_{th}) candidate beam Defined as 1 where the aperture exists and 0 elsewhere
u	Vector	Aperture variable that indicates MLC leaf positions for all candidate beams $u = \begin{bmatrix} u_{b=1\ \alpha=1}^T & u_{b=1\ \alpha=2}^T & \cdots & u_{b=1\ \alpha=n_\alpha}^T & u_{b=2\ \alpha=1}^T & u_{b=2\ \alpha=2}^T & \cdots & u_{b=n_b\ \alpha=n_\alpha}^T \end{bmatrix}^T$
Other data		
S	Set	A set of all feasible candidate beams
$A_{b\alpha}$	Matrix	Fluence to dose transformation matrix for the $(\mathrm{b_{th}}, \alpha_{\mathrm{th}})$ candidate beam
W	Matrix	Diagonal weighting matrix, with weightings for structures of interest as diagonal elements
d_0	Vector	Ideal dose with the prescription dose at the PTV and zero elsewhere
$D_{1_{b\alpha}}$	Matrix	Derivative matrix in the direction parallel to the MLC leaf movement for the $(\mathrm{b_{th}}, \alpha_{\mathrm{th}})$ candidate beam
$D_{2_{b\alpha}}$	Matrix	Derivative matrix in the direction orthogonal to the MLC leaf movement for the $(\mathrm{b_{th}}, \alpha_{\mathrm{th}})$ candidate beam
$P_{b\alpha}$	Scalar	$\mathrm{P}_{\mathrm{b}\alpha}$ is 1 for candidate beams on the selected trajectory from BTS and 0 elsewhere
I_{PTV}	Matrix	Indicator diagonal matrix for PTV, with its diagonal elements equal to 1 for voxels in PTV and 0 elsewhere
$n_{b\alpha}$	Scalar	Number of beamlets with a trajectory that intersects PTV in the $(\mathrm{b_{th}}, \alpha_{\mathrm{th}})$ candidate beam
$G_{b\alpha}$	Scalar	Weightings of the group sparsity term for each feasible candidate beam to compensate for unfair penalization on candidate beams $G_{b\alpha} = \frac{sum(I_{PTV} A_{b\alpha} \vec{1})}{\sqrt{n_{b\alpha}}}$
D_P	Matrix	Derivative matrix in the direction of the selected trajectory, which calculates the difference in MLC leaf positions between adjacent beams on the selected trajectory.

Table 4-15 Notations and data structures used in this study.

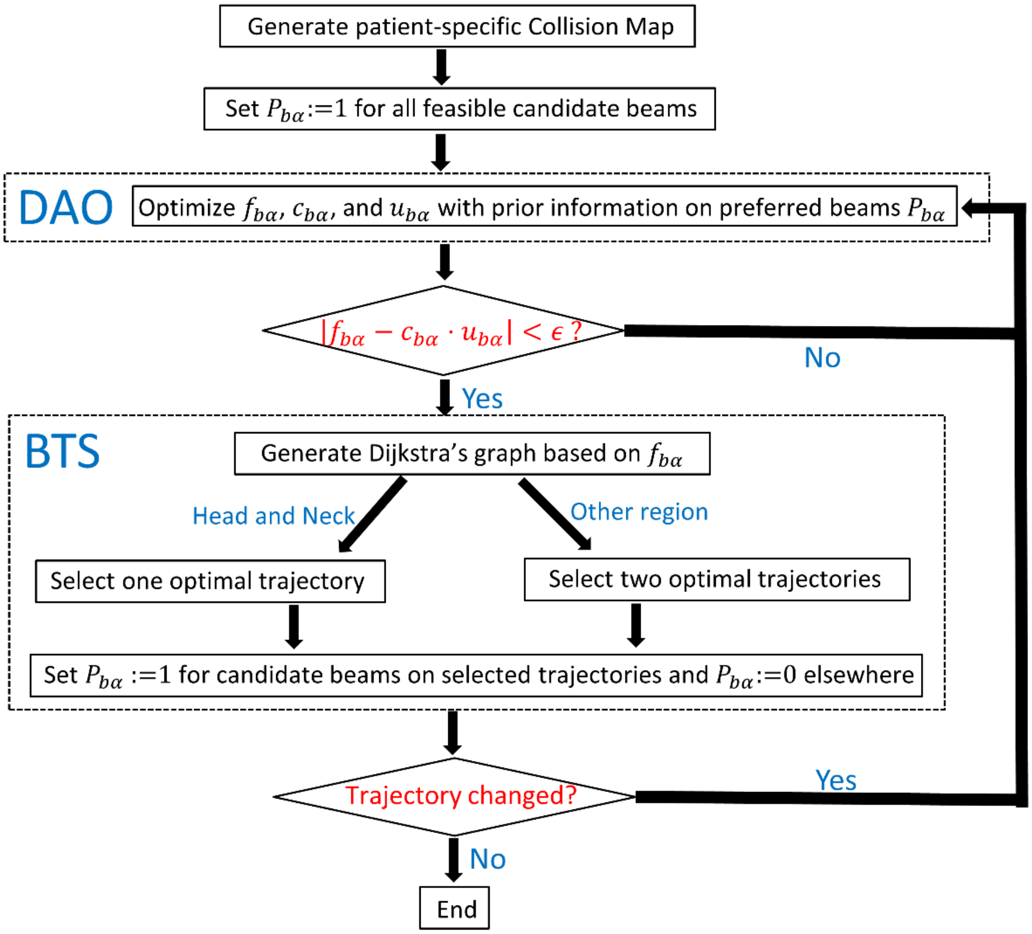


Figure 4-28 Flowchart of 4πVMAT optimization

4.5.2.2 Beam trajectory selection

The BTS is formulated as a travelling salesman problem, where the gantry/couch graph $G = [N, E]$ is defined as a number of nodes N and edges E that connects every two nodes. In this 4πVMAT optimization framework, the gantry/couch graph contains n_α by n_b nodes, representing n_α gantry angles and n_b couch angles, and associated with a node cost $NC(b, \alpha)$ for each candidate beam (b, α) and the edge cost $EC\big((b_1, \alpha_1), (b_2, \alpha_2)\big)$ for the edge that connects (b_1, α_1) and (b_2, α_2). A graph search algorithm is performed on the graph to determine the shortest path from one node to any other nodes. For the patient safety and comfort, this 4πVMAT framework enforces a constant couch rotation direction within each arc and allows the gantry to rotate dynamically. The 4πVMAT arc is represented by a continuous, one-way path on the gantry/couch graph, starting from the first couch angle and ending at the last couch angle, containing only one candidate beam per couch angle.

The node cost $NC(b,\alpha)$ for a feasible candidate beam with couch angle b and gantry angel α depends on $f_{b\alpha}$, the fluence map from DAO&BOO at the current iteration, which is deliverable for each candidate beams, though they may not be connected by a trajectory. The node cost $NC(b,\alpha)$ for an infeasible candidate beam is infinity, which enforces trajectory selection from feasible beams only. The formulation of $NC(b,\alpha)$ is given by

$$Cost_{max} = max\left\{\frac{1}{2}\ \|W(A_{b\alpha}f_{b\alpha} - d_0)\|_2^2\ \middle|\forall(b,\alpha) \in S\right\}$$

$$NC(b,\alpha) = \begin{cases} -log\left(Cost_{max} - \frac{1}{2}\ \|W(A_{b\alpha}f_{b\alpha} - d_0)\|_2^2\right), & if\ (b,\alpha) \in S \\ \infty, & if\ (b,\alpha) \notin S \end{cases}.$$

Equation 4-13

Candidate beams that correspond to dose contribution closer to the ideal dose distribution are assigned with lower node costs. By finding the shortest path on the Dijkstra's graph, the optimal trajectory prefers candidate beams that are more dosimetrically promising.

The edge costs between every two nodes enforce constraints on trajectory selection. The edge cost between candidate beams (b_1,α_1) and (b_2,α_2) is defined as

$$EC\big((b_1,\alpha_1),(b_2,\alpha_2)\big) = \begin{cases} 0 & if\ \ b_2 - b_1 = 1\ \ and\ \ \|\alpha_1 - \alpha_2\| < \alpha_0 \\ \infty & otherwise \end{cases},$$

Equation 4-14

where α_0 is calculated as $\alpha_0 = \omega_g \cdot \tau$, with ω_g representing the maximum angular rotation speed of the gantry and τ the time frame between each control point. Equation (3) ensures that the selected trajectory is a continuous one-way path that contains only one gantry angle per couch angle, and that the gantry angles of adjacent beams on the selected trajectory are within gantry rotation speed limits (controlled by α_0). For simpler cases such as the brain tumor, a single 4πVMAT arc can be adequate to achieve satisfactory dose profile. For more complex patients and body sites where non-coplanar angles are restricted by collision, two arcs may be necessary to substantially improve the dose distribution. In this case, the node cost $NC(b, \alpha)$ is assigned to be infinity for any candidate beams on the first selected trajectory, and then the second trajectory is chosen by finding the shortest path on the updated graph.

4.5.2.3 Algorithm

Here we provide a solution to the optimization framework. The DAO&BOO module was solved using FISTA, and the BTS module was solved using Dijkstra's algorithm. Together the optimizer solves the entire problem iteratively and alternatingly between the two modules. Akin to the previous VMAT optimization algorithms in sections 4.1, 4.2, and 4.4, the DAO&BOO module consists of three submodules, each of which solves the optimization problem in Equation 4-12 with respect to one optimization variable while holding the other two variables constant, and the whole optimization problem is solved in an alternating block fashion.

The edge costs in the BTS problem are in a special form, where all connections between two candidate beams are assigned as infinity except for the path that goes directly to the next couch angle with a moderate gantry angle variation. It ensures that the selected trajectory starts from the first couch angle and ends at the last couch angle, with only one candidate beam selected for each couch angle. This setup not only guarantees a practicable trajectory for 4πVMAT delivery that addresses patient safety and comfort, but also makes the travelling salesman problem more straightforward and computationally inexpensive to solve. This special graph optimization problem is solved using a simplified Dijkstra's algorithm[154,164] in Table 4-16 α_{short} is a vector with n_b elements denoting the candidate beams on the selected trajectory, with couch angle b and gantry angle $\alpha_{short}(b)$. For lung and prostate patient, a secondary trajectory α_{short2} is obtained on top of α_{short}. Once the trajectories are chosen, $P_{b\alpha}$ is assigned to be 1 for any selected candidate beams, and 0 elsewhere, and is utilized in DAO&BOO to assign different group sparsity penalties for selected and unselected candidate beams.

```
Simplified Dijkstra's algorithm
Initialize the graph with each node $G(b,\alpha) \coloneqq \infty$
$G(1,\alpha) \coloneqq NC(1,\alpha)$
for $b = 1,2,\dots,n_b - 1$
    for $\alpha = 1,2,\cdots,n_\alpha$
        $G(b+1,\alpha) \coloneqq \underset{\alpha'}{min}\ G(b,\alpha') + NC(b+1,\alpha) + EC\big((b,\alpha'),(b+1,\alpha)\big)$
        $\alpha_t(b+1,\alpha) \coloneqq \underset{\alpha'}{argmin}\ G(b,\alpha') + NC(b+1,\alpha) + EC\big((b,\alpha'),(b+1,\alpha)\big)$
    end for
end for
$\alpha_{short}(n_b) \coloneqq \underset{\alpha'}{argmin}\ G(n_b,\alpha')$
for $b = n_b - 1, n_b - 2, \dots, 1$
    $\alpha_{short}(b) \coloneqq \alpha_{short}\big(b+1,\alpha_1(b+1)\big)$
end for
if planning for lung and prostate cancer patient
    $NC\big(b,\alpha_{short}(b)\big) \coloneqq \infty$
    $G_2(1,\alpha) \coloneqq NC(1,\alpha)$
    for $b = 1,2,\cdots,n_b$
      for $\alpha = 1,2,\cdots,n_\alpha$
          $G_2(b+1,\alpha) \coloneqq \underset{\alpha'}{min}\ G_2(b,\alpha') + NC(b+1,\alpha) + EC\big((b,\alpha'),(b+1,\alpha)\big)$
          $\alpha_{t2}(b+1,\alpha) \coloneqq \underset{\alpha'}{argmin}\ G_2(b,\alpha') + NC(b+1,\alpha) + EC\big((b,\alpha'),(b+1,\alpha)\big)$
      end for
    end for
    $\alpha_{short2}(n_b) \coloneqq \underset{\alpha'}{argmin}\ G_2(n_b,\alpha')$
    for $b = n_b - 1, n_b - 2, \cdots, 1$
      $\alpha_{short2}(b) \coloneqq \alpha_{t2}\big(b+1,\alpha_{short2}(b+1)\big)$
    end for
End
```

Table 4-16 Pseudocode for Simplified Dijkstra's algorithm

4.5.2.4 Implementation Details

The feasible beam set is based on the collision model from our previous study[133], which provides an individualized collision prediction model for the purpose of non-coplanar beam delivery in IMRT. In the 4πVMAT study, all candidate beams that are predicted to be undeliverable isocentrically are excluded, and a 10-degree margin is added to both the upper and lower bounds of the infeasible gantry angle range for each couch angle to avoid couch-gantry-patient collision during the gantry/couch rotation between feasible beams.

Parameter tuning can be tricky in this 4πVMAT framework since there are a number of hyper-parameters in addition to the regular structure weightings that help ensure the deliverability of the 4πVMAT plan. Nevertheless, the optimization results are insensitive to most of the hyper-parameters that control the deliverability. Once a suitable set of these parameters are found, our experiments show robust performance even if the structure weightings are changed or applied to different patient cases with comparable sizes of PTV and OARs.

To achieve a satisfactory local minimum in the data fidelity term, we set up some heuristics for parameters $\gamma_1 - \gamma_4$. All of them increase as the optimization progresses, which allows the optimization to focus on the convex dose fidelity term and ensure a good local minimum with high dosimetric quality at the initial stages of the optimization. During the optimization, the single segment term and group sparsity term are gradually emphasized to ensure plan deliverability of the DAO result. γ_1 and γ_2 were updated when each round of three submodules in DAO is completed

$$\gamma_1 = \gamma_{10} \cdot \left(1 + (1000)^{\frac{i}{N}}\right) \qquad \gamma_2 = \gamma_{20} \cdot \left(1 + (1000)^{\frac{i}{N}}\right),$$

where i is the number of rounds from submodule 1 to submodule 3 within the DAO module, N is the total number of rounds within the DAO module, which we chose to be 6 in this study empirically. γ_{10} and γ_{20} are the initial values of γ_1 and γ_2 respectively. γ_3 and γ_4 enforces the number of active candidate beams to be within certain ranges, which is designed to decrease with the iteration of BTS process. In submodule 1, where the group sparsity term controls the sparsity level of $f_{b\alpha}$, the number of active beams is evaluated every 20 iterations. γ_3 and

γ_4 are automatically increased or decreased by 20% if there are too many or insufficient active beams.

All the hyperparameters were tuned based on the optimization performance. For example, the λ_1 and λ_2 would be increased if the fluence map was too rough or decreased if it was overly smooth. We also increase the values of γ_{10} and γ_{20} if there are multiple segments in one candidate beam. The range of these hyperparameters in our study are listed in Table 2.

λ_1	λ_2	γ_{10}	γ_{20}	γ_3	g_1	g_2	g_3
10-1000	0.2-20	0.05-50	0.05-50	50-2000	$10^{-5}-10^{-3}$	$10^{-5}-10^{-3}$	$10^{-5}-10^{-3}$

Table 4-17 The range of the hyperparameters used in this study

4.5.2.5 Evaluation

	Number of feasible beams	**Prescription Dose (Gy/fx)**	**PTV Volume (cc)**
GBM1	1824	25/5	6.23
GBM2			2.34
GBM3			0.77
LNG1	1174	50/4	139
LNG2			10.2
LNG3			116
PRT1	1200	40/5	111
PRT2			127
PRT3			85

Table 4-18 Number of feasible beams, prescription doses for each fraction, and PTV volumes for all patients

The feasibility of the optimization algorithm was tested on three glioblastoma multiforme retreatment patients (GBM), three lung cancer patients (LNG), and three prostate cancer patients (PRT). The GBM retreatment planning followed an internal protocol to minimize dose to previously irradiated critical organs [120]. The centrally located lung and the prostate stereotactic body radiotherapy (SBRT) plans follow [165] and [166], respectively. Table 3 summarizes the number of feasible beams, the prescription doses, and PTV volumes for all patients.

Using a convolution/superposition code with a 6 MV x-ray polyenergetic kernel as described in our previous publications[100], the beamlet dose was calculated for all feasible beams among 2400 candidate beams in 4πVMAT. Since the gantry rotation angles are more likely to be collision-free when the couch stays close to central position, the couch angles are more densely sampled for couch angles ranging from 15º to 0º and from 360º to 345º, following the International Electrotechnical Commission (IEC 61217) convention. The 2400 candidate beams correspond to 30 gantry angles with 12-degree spacing from 0 to 359 degrees, and 80 couch angles, separated by 1 degree for the central 30-degree range and 3 degrees elsewhere, from 270 to 90 degrees. For 2πVMAT, the beamlet dose was calculated for 80 gantry angles with 4.5 degrees of separation for each arc, from 0 to 359 degrees. The beamlet resolution was 0.5×0.5 cm^2, and the dose array resolution was $0.25\times0.25\times0.25$ cm^3.

All arcs contain 80 control points in this study. The 4πVMAT plan contains one arc for the GBM patient, 80 control points in total, which is compared against a 2πVMAT plan with two coplanar arcs with collimator angles at 45º and 135º, 160 control points in total. For the lung cancer patient and prostate cancer patient, the 4πVMAT plan contains two arcs with 160 control points in total, while the 2πVMAT contains three arcs with collimator angles at 30º, 90º, and 150º, 240 control points in total. The 2πVMAT and 4πVMAT utilize the same DAO model except that the 2πVMAT has a predefined coplanar trajectory. The traveling time between each beam is 2s, allowing 12º of gantry rotation.

PTV statistics including PTV D95, D98, D99, D2, (defined as the dose which is received by at least 95%, 98%, 99%, and 2% of the volume, respectively), and PTV homogeneity (defined as $\frac{\mathrm{D95}}{\mathrm{D5}}$) were evaluated. For OAR, the Dmax and Dmean were assessed. The maximum dose is defined as D2, recommended by the ICRU-83 report [101]. The dose conformity, defined as the ratio between the PTV volume receiving 100% or more of the prescription dose and the PTV volume, and the Integral Dose defined as the volume integral of the dose deposited in the patient, were also obtained. The R50, defined as the 50% isodose volume divided by the target volume, was evaluated to quantify the amount of high-dose spillage in the patient body.

4.5.3 Results

Figure 4-29 shows the fluence map of the GBM #1 patient as an example of the optimization result on the Gantry/Couch graph. The red regions on the Gantry/Couch graph indicate the candidate beams that cause collision. Notice that only those candidate beams that are on the selected trajectory have nonzero fluence weights. To avoid couch back and forth motion in 4πVMAT delivery, the couch rotates mono-directionally from 90º to 270º, and the gantry can rotate in both directions.

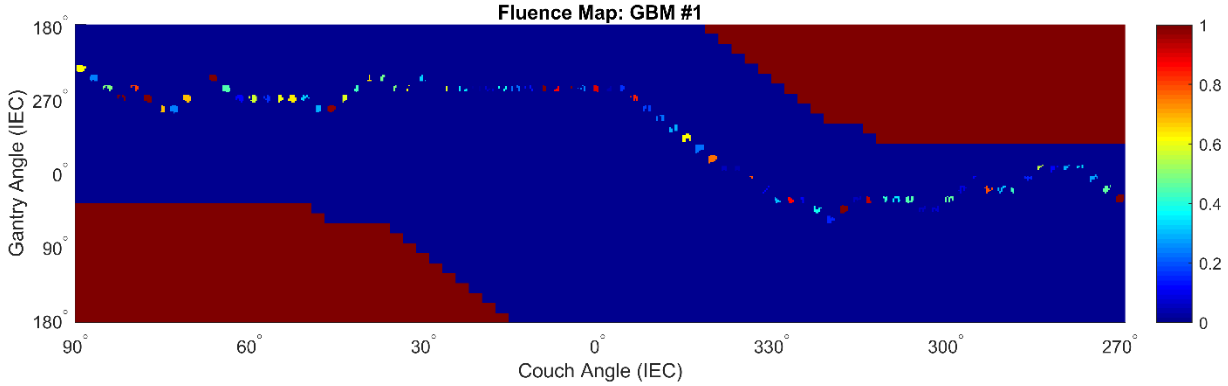


Figure 4-29 Normalized fluence map on the Gantry/Couch graph (GBM #1). Only those candidate beams that are on the selected trajectory have nonzero fluence weights. The couch rotates from 90º to 270º and the gantry rotates accordingly. MLC leaf direction is vertical for this diagram. The red regions denote the gantry/couch angles that cause collision. The relative intensities of apertures are indicated by the colorbar.

Figure 4-30 shows the selected beam angles on the Gantry/Couch graph with the corresponding 3D view for all patients. The 4πVMAT algorithm optimizes for all candidate beams simultaneously and finds one or two optimal trajectories that go through the safe region, from the first couch angle to the last couch angle, indicated by the green blocks, affording efficient delivery of non-coplanar arc. The maximum allowance of the gantry rotation is 12º between two adjacent beams. With the gantry rotation speed limit at 6º per second, 2 seconds is sufficient for delivery at one control point. For the GBM patient, the 4πVMAT plan takes around 3 minutes in estimation to deliver the single arc 4πVMAT plan, and for the lung cancer patient and prostate cancer patient, the calculated time is approximately 5 minutes to deliver the two arcs 4πVMAT plan.

Figure 4-31 shows the dose distribution of 4πVMAT and 2πVMAT for all patients. By using non-coplanar beams, 4πVMAT has the flexibility to distribute the dose in any non-colliding direction within the 4π spherical space, depending on the benefits of OAR sparing and target coverage. Greater separation of these non-coplanar beams also reduces the high dose spillage in 4πVMAT. In the GBM cases where the anatomy is relatively simple, single-arc 4πVMAT plans were able to significantly avoid dose spillage to the brainstem as well as other critical structures, such as the chiasm and right optic nerve in the GBM #1 case, and the left cochlea in the GBM #2. In all the LNG cases, by utilizing beams with greater separation, the two-arc 4πVMAT plans substantially reduced the high dose spillage compared with the three-arc 2πVMAT plan, including those to the critical organs, such as the chest wall, proximal bronchus, and spinal cord in the LNG #1, the heart in LNG #2, and the aorta and pulmonary vessel in LNG #3. For all three prostate cases, the two-arc 4πVMAT plans resulted in a more desirable asymmetric dose in the anterior/posterior direction to substantially better spare the rectum as compared with the three-arc 2πVMAT plan, while achieving similar bladder high dose volumes. In all cases, by utilizing these OAR-sparing angles, 4πVMAT achieved consistently better OARs sparing at the same time maintaining or improving PTV coverage.

Figure 4-32 shows the zoomed-in view of the DVHs of 4πVMAT and 2πVMAT for all OARs, Figure 4-33, Figure 4-34, and Figure 4-35 show the complete DVHs. The 4πVMAT was able to markedly reduce dose to OARs while achieving comparable or better PTV statistics across all patients, especially for the dose limiting organs, such as the brainstem in the GBM #2 and GBM #3, the proximal bronchus in all three LNG patients, the major vessels in LNG #2, and LNG #3, and the seminal vesicle and the rectum in all PRT patients. In the GBM #3 case, the one-arc 4πVMAT plan reduced the maximum doses to the brainstem by 8.1 Gy (64.8%) compared with the two-arc 2πVMAT plan. In the LNG #1 case and the LNG #2 case, the two-arc 4πVMAT plan reduced the dose to the proximal bronchus and the major vessels by 16.3 Gy (41.5%) and 19.83 Gy (55.5%) compared with the three-arc 2πVMAT plan. Across all LNG case, the 4πVMAT plan substantially reduced the hot spots in the critical structures that were covered by high dose up to 30Gy.

The quantitative statistics for the PTV and OARs are shown in Table 4-19 and Table 4-20. The PTV statistics, Dose Conformity, and Integral Dose are comparable across 4πVMAT and 2πVMAT, but the 4πVMAT is able to substantially reduce R50, indicating a remarkable improvement in dose compactness that is consistent with previously reported 4π IMRT-VMAT comparison. On average, the 4πVMAT plan reduced the OARs max and mean doses by 9.63% and 3.08% of the prescription dose. The single largest sparing in OARs max and mean dose is up to 39.7% and 14.9% of the prescription dose.

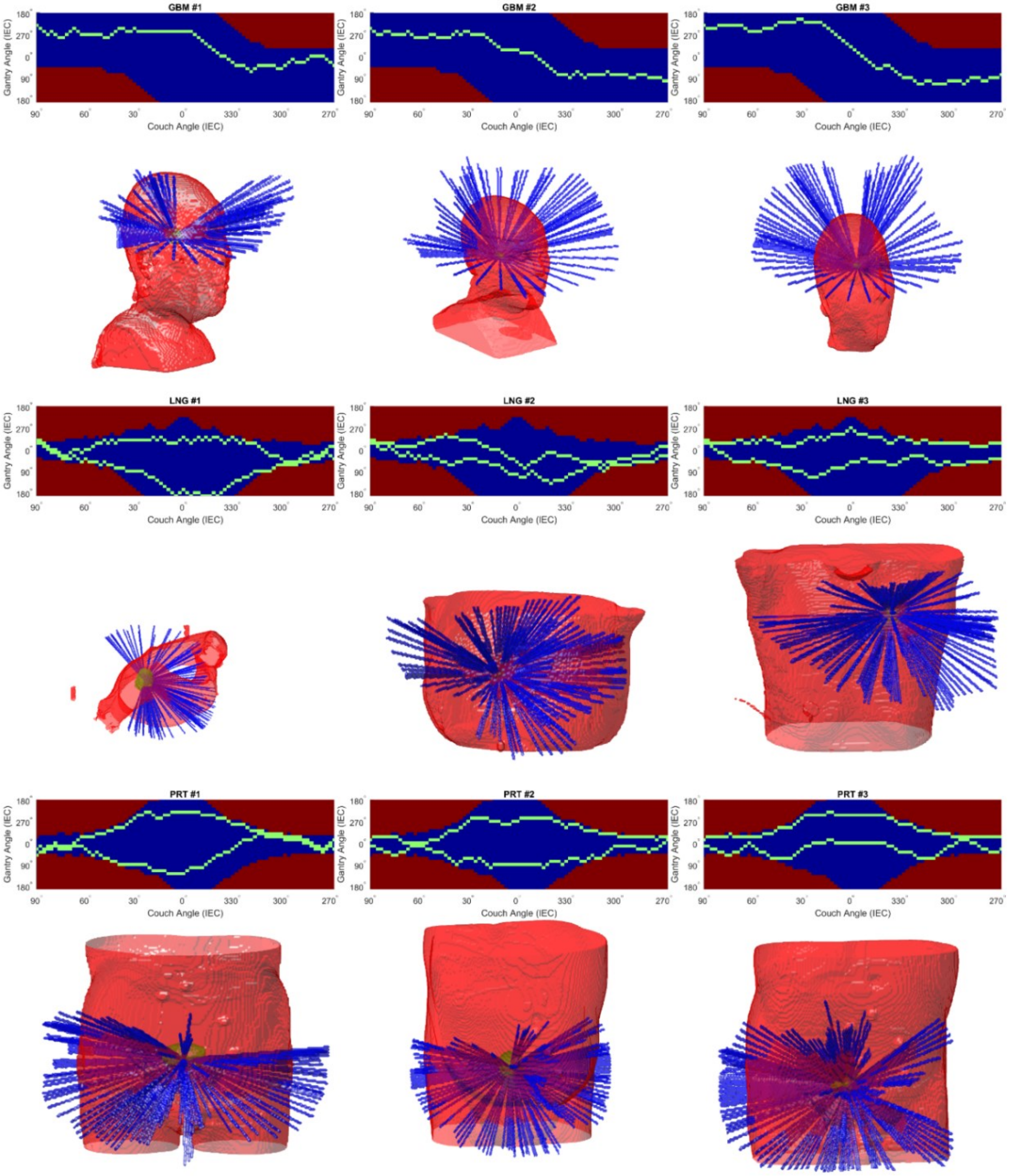


Figure 4-30 The selected beam angles on the Gantry/Couch graph with corresponding 3D view for all patients. The red regions on the Gantry/Couch graph indicate the candidate beams that cause collision. The green blocks show the selected trajectories.

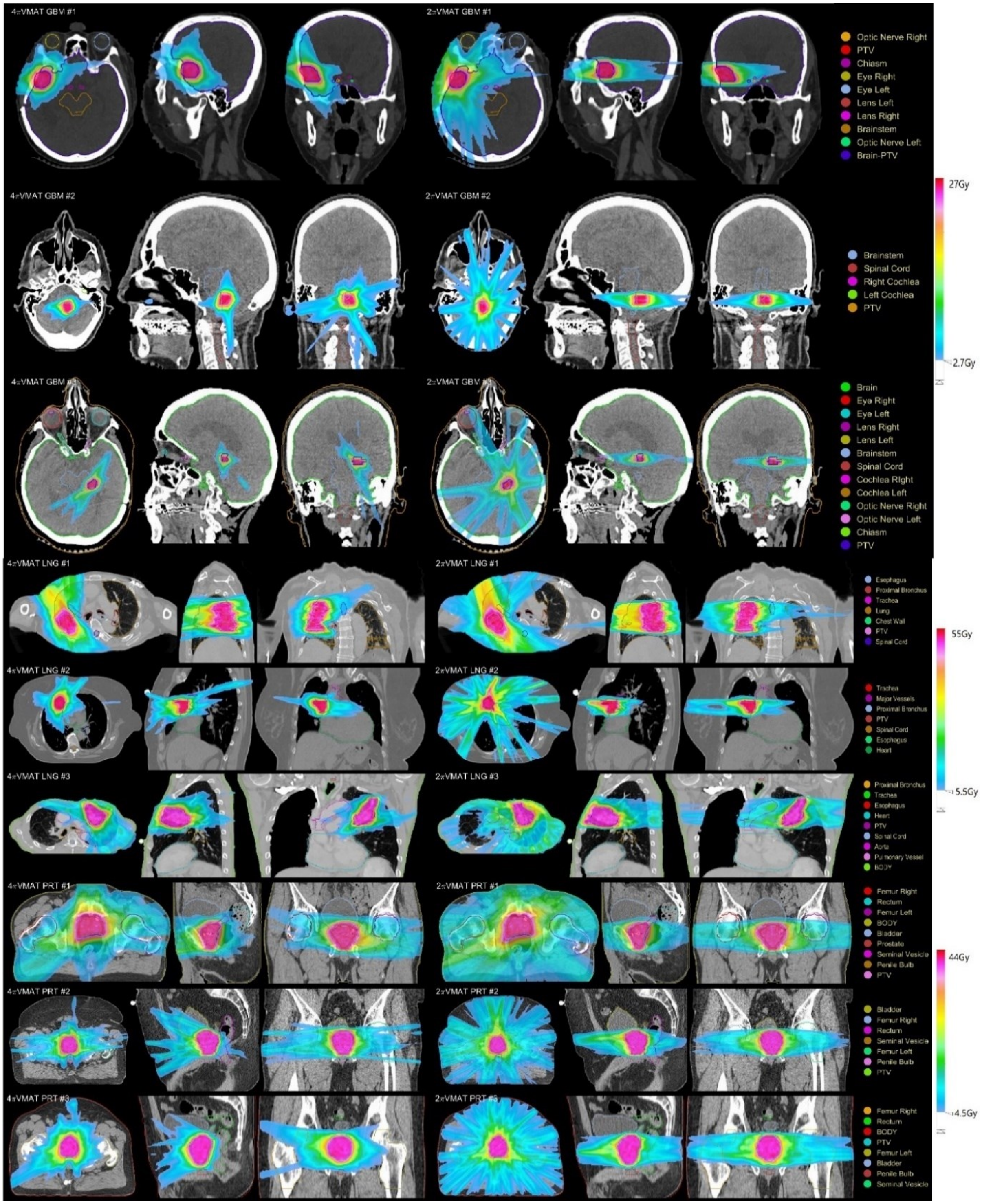


Figure 4-31 Isodose colorwash comparison for all patients.

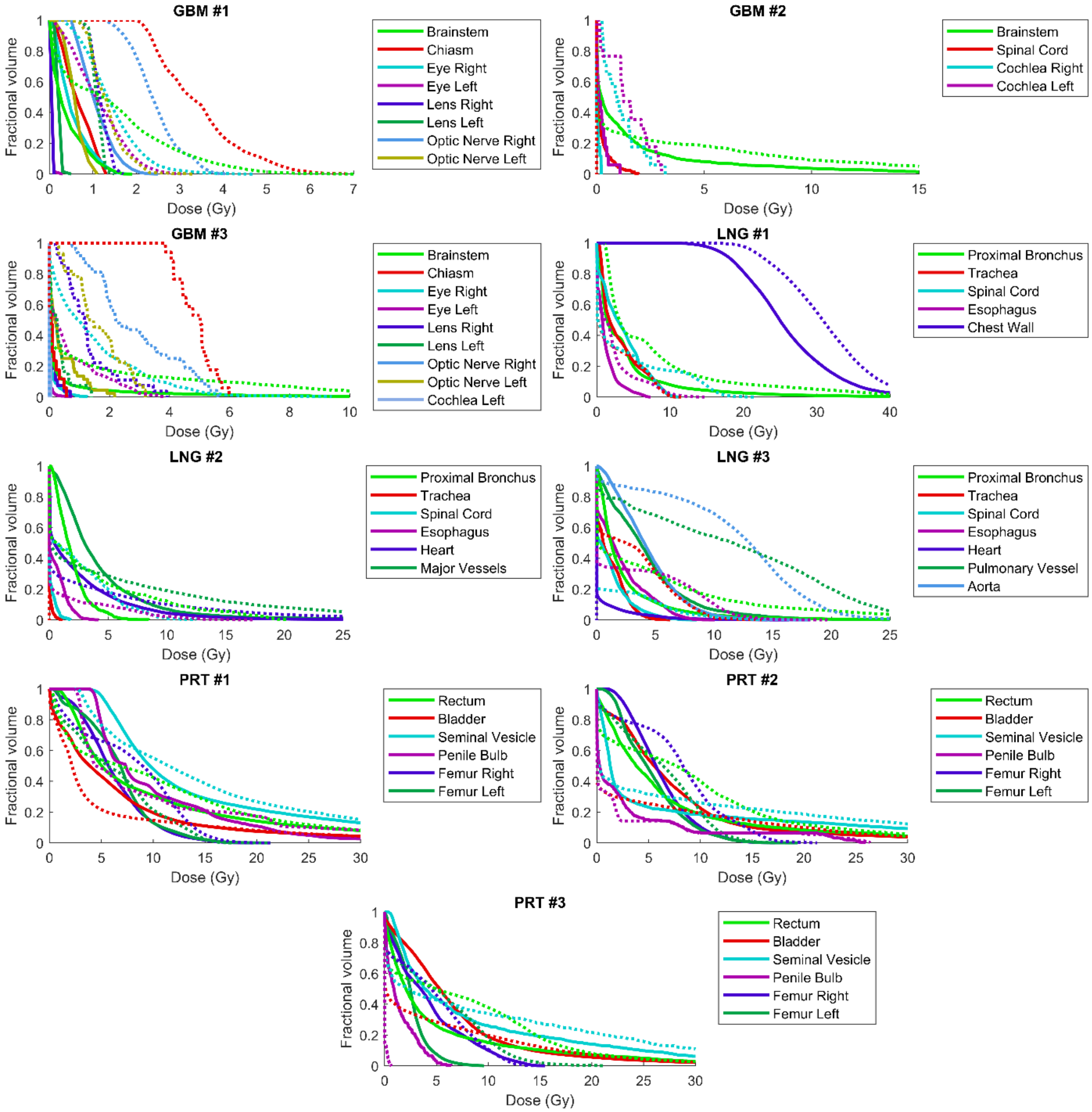


Figure 4-32 DVH comparison of the 4πVMAT (solid) and the 2πVMAT (dotted) for all OARs of the patients.

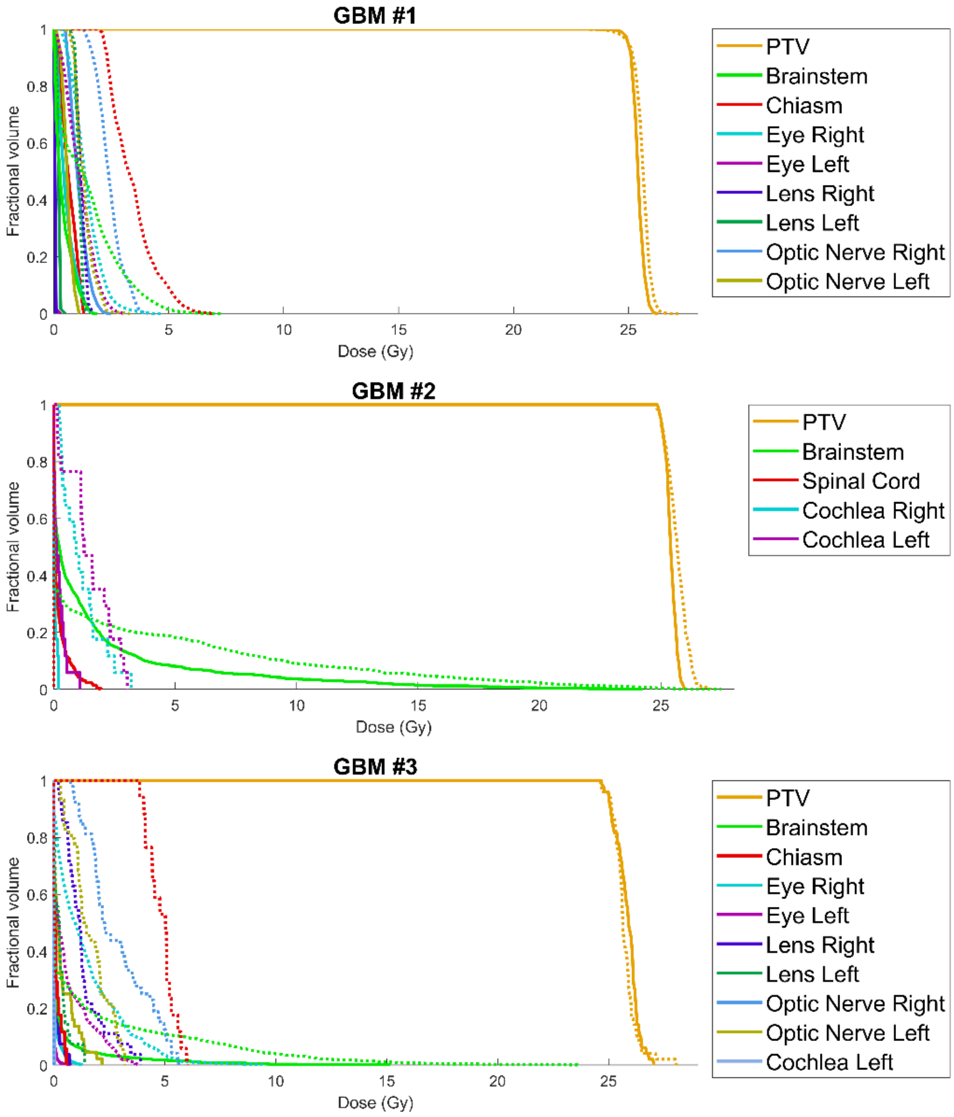


Figure 4-33 DVH for GBM patients. The solid lines are for 4πVMAT and the dotted lines are for 2πVMAT. D95 is normalized to the prescription dose.

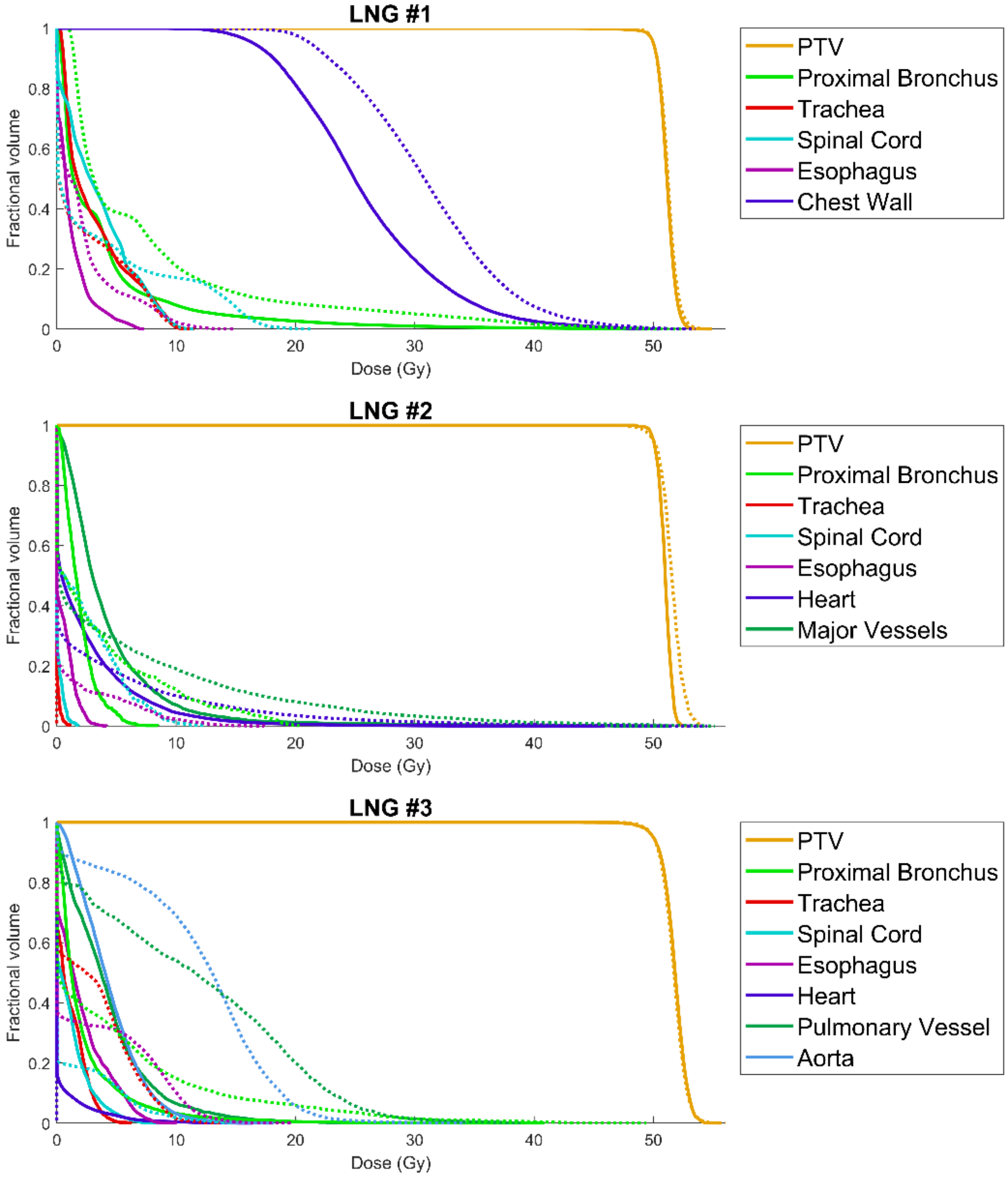


Figure 4-34 DVH for LNG patients. The solid lines are for 4πVMAT and the dotted lines are for 2πVMAT. D95 is normalized to the prescription dose.

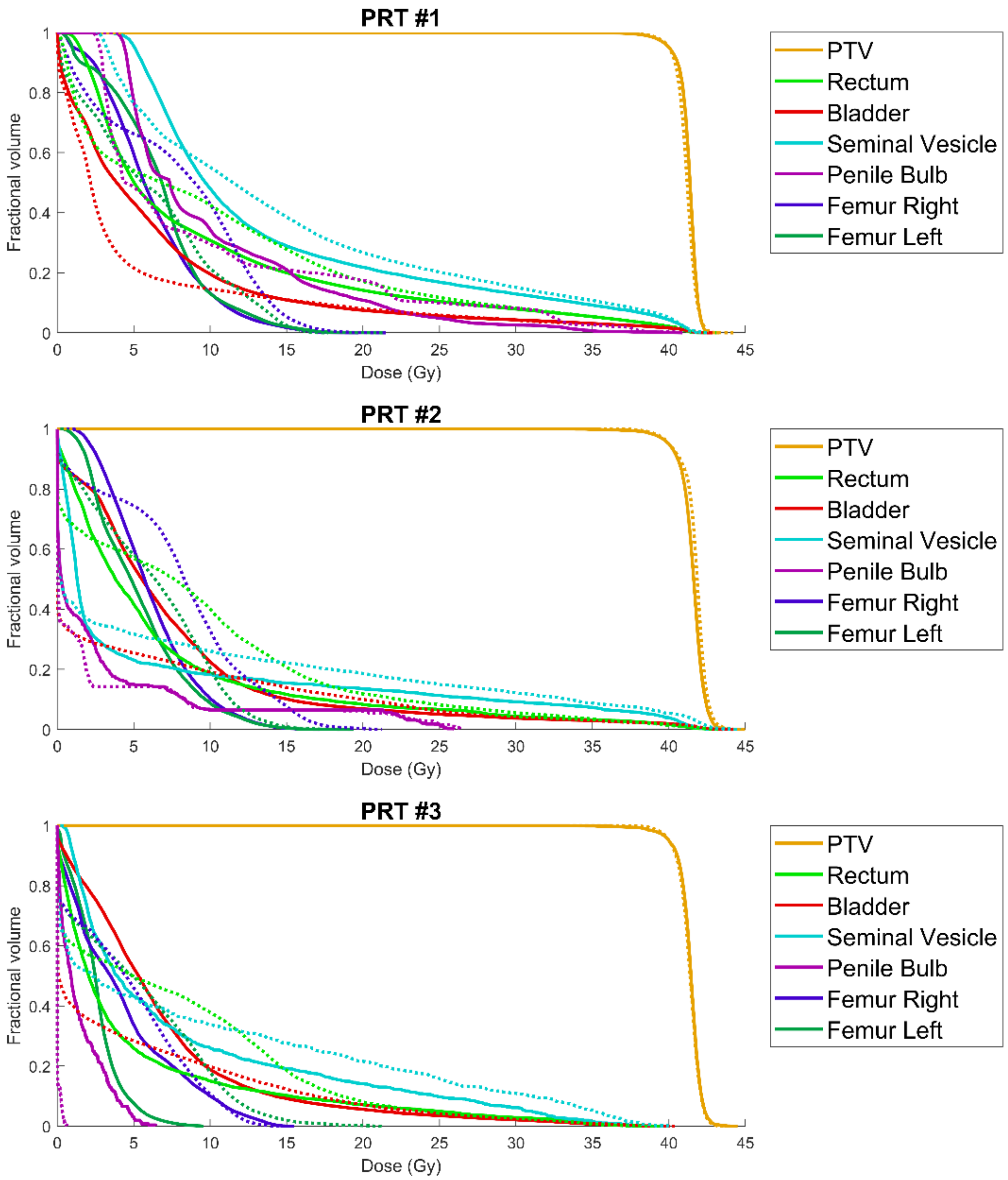


Figure 4-35 DVH for PRT patients. The solid lines are for 4πVMAT and the dotted lines are for 2πVMAT. D95 is normalized to the prescription dose.

Patient Case	PTV Statistics						Dose Conformity		R50		Integral Dose ($10^3\ Gy \cdot cm^3$)	
	Homogeneity		D95	D98	D99	Dmax						
	4π VMAT	2π VMAT	4πVMAT – 2πVMAT (Gy)				4π VMAT	2π VMAT	4π VMAT	2π VMAT	4π VMAT	2π VMAT
GBM1	0.97	0.96	-0.01	0.16	0.31	-0.11	0.95	0.95	3.84	5.02	2.34	2.06
GBM2	0.97	0.95	0.02	0.03	0.09	-0.56	0.95	0.94	3.52	5.20	2.13	1.49
GBM3	0.94	0.94	0.00	0.05	0.04	-1.06	0.94	0.95	4.74	6.84	0.85	0.87
LNG1	0.96	0.95	0.00	-0.05	-0.20	-0.33	0.95	0.95	5.36	6.98	74.60	83.13
LNG2	0.97	0.94	-0.01	0.33	0.82	-1.50	0.95	0.95	3.77	6.77	15.17	15.99
LNG3	0.94	0.94	-0.01	-0.08	-0.27	-0.04	0.95	0.95	2.37	2.87	36.01	43.86
PRT1	0.95	0.95	0.00	-0.16	-0.26	-0.06	0.95	0.95	3.25	3.72	47.12	47.26
PRT2	0.94	0.93	0.00	-0.20	-0.28	-0.21	0.95	0.95	2.19	2.51	57.24	53.48
PRT3	0.94	0.95	-0.01	-0.27	-0.66	0.11	0.95	0.95	1.45	1.62	34.42	31.28

Table 4-19 PTV statistics for all patients

OAR dose sparing: 2πVMAT - 4πVMAT (Gy)				
	Dmax		Dmean	
	Largest Value	Average Value	Largest Value	Average Value
GBM1	3.21 Brainstem	1.88	2.33 R Opt Nrv	1.17
GBM2	6.68 Brainstem	2.51	1.25 L Cochlea	0.77
GBM3	8.10 Brainstem	3.63	2.85 R Opt Nrv	1.49
LNG1	16.3 Proximal Bronchus	6.39	5.29 Chest wall	1.81
LNG2	19.83 Major Vessels	10.11	2.13 Spinal Cord	0.88
LNG3	15.48 Proximal Bronchus	7.19	7.46 Aorta	2.77
PRT1	4.20 Penile Bulb	1.09	1.72 R Femur	0.14
PRT2	3.80 R Femur	1.09	2.30 Rectum	0.56
PRT3	8.22 L Femur	0.70	2.70 L Femur	0.53

Table 4-20 OAR mean and maximum dose sparing differences between 4πVMAT and 2πVMAT for all patients. The columns labelled “Largest Values” represents the largest amount of dose sparing difference achieved among all OARs, and the corresponding OAR. “Average values” represents the average sparing difference among all OARs

4.5.4 Discussion

Our previous 2πVMAT method solves the VMAT optimization problem by considering all beams simultaneously instead of progressive sampling, making it possible to incorporate BOO into non-coplanar VMAT. The integrated 4πVMAT optimization framework further develops the 2πVMAT method by simultaneously solving both the DAO and BTS problems. By expanding the solution to the non-coplanar space, 4πVMAT was able to greatly improve the dose compactness as indicated by R50, reduce dose spillage to OARs and subsequently achieve better dosimetry than the 2πVMAT with more arcs. Considering that the non-progressive sampling 2πVMAT was already superior to the existing progressive sampling VMAT algorithm that is widely employed in clinics, 4πVMAT holds the strong promise of substantially improving state-of-the-art radiotherapy without sacrificing delivery efficiency.

4πVMAT takes an alternating approach between DAO&BOO and BTS. From the dosimetric point of view, the DAO&BOO achieves high plan quality by activating a set of dosimetrically promising beams, and the BTS finds the optimal trajectory that tends to include most of these active beams by minimizing the selected node costs. On the other hand, to address the mechanical constraints, the regularization in the DAO&BOO enforces the plan deliverability, and the BTS meets the delivery/treatment time constraint and selects only the trajectories that does not require a substantial gantry rotation between adjacent beams by minimizing the edge cost. The selected trajectory is then translated into DAO&BOO to further explore the dose domain through a weighted group sparsity term, where off-trajectory candidate beams are penalized more heavily. In the DAO&BOO process, promising off-trajectory beams could still be turned on to provide better dose distribution, and on-trajectory beams might be rejected if they are dosimetrically undesirable. By alternatingly optimizing between the

DAO&BOO and BTS, the algorithm makes an integrated final decision based on both the dosimetry and mechanical constraints. With the gantry rotation speed limit at 6º per second, 2 seconds is sufficient for delivery at one control point of the current 4πVMAT plan, where the largest gantry rotation angle is 12 degrees between adjacent beams. A more rapid delivery may be achieved with faster gantry rotation.

Compared with the previous non-coplanar VMAT study with user-defined heuristic trajectories [118,158,159] and trajectory optimization techniques using geometric information [113,160,161], this 4πVMAT optimization framework is able to thoroughly and automatically search the entire non-coplanar trajectory space for various patient anatomies and dosimetric requirements. Compared with other non-coplanar VMAT optimization algorithms using optimized static IMRT beams as anchoring nodes for non-coplanar arcs that may not be dosimetrically optimal [162,163], our 4πVMAT integrates BTS into VMAT optimization and encourages the fluence map to be developed natively into 4πVMAT arcs. Compare with the current commercial solution HyperArc[167], which is restricted to using a few predefined trajectory templates and is currently only applied to the brain treatment, the proposed 4πVMAT framework fully explores the 4π space with a site-specific collision model, and hence holds the promise of substantially improving the dose profile and being more broadly applicable to other body sites.

4πVMAT in its current form is computationally intensive due to the optimization problem size and alternating optimization between DAO and BTS despite our effort to accelerate the computation. Even with the accelerated algorithm FISTA, the optimization run time on a single desktop using MATLAB implementation still took from 1 hour for the GBM case to 9 hours for the LNG case and PRT case. To further speed up optimization for clinical implementation, aside from switching to a higher performance language such as C, the most computationally expensive matrix multiplication step in FISTA is parallelizable and can therefore be moved to graphic processing units (GPUs) for higher computation efficiency. The current 4πVMAT framework was evaluated with 12 degrees of gantry angle spacing to reduce the computation cost. With future acceleration techniques, we may be able to perform the optimization with a denser sampling of the 4π space.

Another limitation of the 4πVMAT optimization framework is its complexity in tuning the hyperparameters, compounded with its heavy computation costs, can be time consuming and laborious. One way to get around with this problem is to simplify the dataset when tuning the hyperparameters, either by downsampling the structure of interest or truncating the dose calculation matrix. From our experience, the algorithm is insensitive to this procedure, hence the same set of hyperparameters apply to the full sampled data.

4.5.5 Summary

The non-coplanar 4πVMAT significantly expands the searching space and improved dose compactness and OAR sparing compared with the coplanar VMAT (2πVMAT) with more arcs.

4.6 ROAD: ROtational direct Aperture optimization with a Decoupled ring-collimator for FLASH radiotherapy

4.6.1 Introduction

The improvement in physical radiation dose conformity via technological evolutions, including IMRT[24,168], VMAT[22,169], and 4π non-coplanar radiation therapy [39,107,120], will plateau. Particularly for normal tissues abutting the tumor, one must look beyond the physical dose for the next leap in the therapeutic ratio. Emerging evidence suggests that ultra-fast radiation delivery (dose rate > 40 Gy/s), termed FLASH radiotherapy, may lead to such a leap[170]. Compared to conventional radiotherapy with dose rate ~ 0.1 Gy/s, FLASH radiotherapy markedly reduced the normal tissue toxicity without compromising tumor response [171–174]. The FLASH effects have been consistently observed across different animal species, including mice[171], cat[172], zebrafish [175], and pig [172], in various tumor models including lung [171], breast [171], and brain [171,176,177], using different modalities including electron [172,178], X-ray [179], and proton [175]. Moreover, the FLASH effect was recently demonstrated in a human study [180].

Nevertheless, there are significant technical challenges to achieving the ~500x greater dose rate for FLASH in human patients. The electron dose rates using existing linacs may be high enough, but the achievable energies are inadequate for most non-superficial tumors. Certain proton systems can be modified to achieve the high dose rate but only in the dosimetrically inferior transmissive mode, due to the non-negligible time required to switch between energy layers in the proton scanning spot mode.

Therefore, we are motivated to investigate an X-ray system for FLASH therapy with two essential components: a high output X-ray linac and the dose conformity comparable to state-of-the-art IMRT. It is believed that the high dose rate can be achieved with existing accelerator technology. Conventional medical linacs produce average electron beam powers of around 1 kW, while higher-power industrial and research accelerators routinely achieve powers on the order of 100 kW[181]. Assuming a dose conversion factor of 7.5 Gy/s/mA at 10 MeV[182], the beam power needed to achieve 1 Gy/s is 80 kW. Furthermore, we note that the ~100 kW industrial accelerators that are in operation around the world are designed to work continuously in a factory setting with very little downtime. In the case of FLASH, the beam will only be on for a brief time, on the order of 1 s, which significantly relaxes the requirements on, e.g., power supplies and cooling systems. While we do not claim that such an accelerator will be easy to produce, it is feasible with current technology.

Multiple beam angles and effective intensity modulation are necessary for good X-ray dose conformity. Existing delivery platforms are evidently incompatible with the FLASH requirement. To avoid the slow mechanical movements of the C-arm gantry and the multi-leaf collimator (MLC), Maxim et al. [183] proposed PHASER using high-intensity, pluridirectional MV X-ray beams and scanning electron pencil beams for IMRT. However, the success of PHASER depends on multiple groundbreaking technologies to be developed, adding significant uncertainties, a long development cycle, and high cost to the product.

Alternatively, to achieve ultrafast dose delivery and intensity modulation with X-rays, we propose to *modify* the existing method for delivering VMAT to enable FLASH delivery. We term the novel delivery method <u>RO</u>tational direct <u>A</u>perture optimization with a <u>D</u>ecoupled ring-collimator (ROAD), which employs a fast-rotating slip ring gantry, and a decoupled MLC-ring with many pre-shaped apertures for fast access to multiple beam entry angles and complex dose modulation.

4.6.2 Methods

4.6.2.1 <u>Hardware design of ROAD</u>

The standard VMAT treatment delivers radiation continuously from a rotating gantry head that encloses a radiation source, jaws, and an MLC module (Figure 4-36A). The MLC leaves move while the gantry rotates to form different aperture shapes at different beam angles. Due to the limited MLC leaf speed (typically <5cm/s at the isocenter distance), a constraint is imposed on the difference between adjacent apertures for practical VMAT delivery time. Reducing the delivery time to sub-second for FLASH would require faster source rotation and a different way to provide intensity modulation.

Figure 4-36B shows the proposed ROAD design. The fast-rotating source is achieved by mounting the linac on a slip-ring gantry with a speed of 1 rotation per second (rps). The linac, which is currently under development, has the following preliminary technical specifications: pulse length 83 μs, repetition rate 300 Hz, dose per pulse 2 Gy at the isocenter, energy 10 MV. To circumvent the mechanical limitations of a single MLC module, ROAD uses 75 MLC modules mounted on a separate ring that is either static or counter-rotating at 1 rps. Different from conventional VMAT, where the MLC leaves are always aligned with the X-ray target, in the decoupled configuration, the individual MLC modules are focused at 75 equally spaced points on the target ring. With an accurate measurement of the linac/MLC angles, the pulsed linac is triggered only when the target is aligned with one of the MLCs to eliminate the undesirable geometrical penumbra. Intensity modulation is then achieved by optimizing a single MLC aperture for each beam. Because apertures can be shaped *before* treatment, the MLC mechanical speed limitation is circumvented.

Figure 4-36C and Figure 4-36D illustrate the ROAD design assuming the following physical and mechanical parameters. The source to isocenter distance (SID) is 100 cm, the distance between the MLC and the isocenter is 70 cm, the MLC thickness is 10cm, and the beamlet resolution is 5mm at the isocenter plane (the MLC physical width is 1.5mm). A total of 75 identical MLC modules are installed on the ring-collimator, with 60 MLC leaves in each MLC module (30 leaves on each side). The beam FOV is 15cm in-plane and 20cm cross-plane. The gantry rotation speed is 1rps, and the MLC leaf traveling speed at the isocenter plane is

5cm/s. With static ring-collimator in one full ring-source rotation, 75 equiangular beams can be utilized (ROAD-75) (Figure 4-36C). By counter-rotating ring-collimator at 1 rps, 150 equiangular beams can be achieved (ROAD-150) (Figure 4-36D). We assume a total delivery time of 1s for both ROAD-75 and ROAD-150. In ROAD-150, the same MLC modules are used twice separated by 0.5 seconds, allowing up to 2.5cm leaf motion at the isocenter plane in the transition to form new apertures.

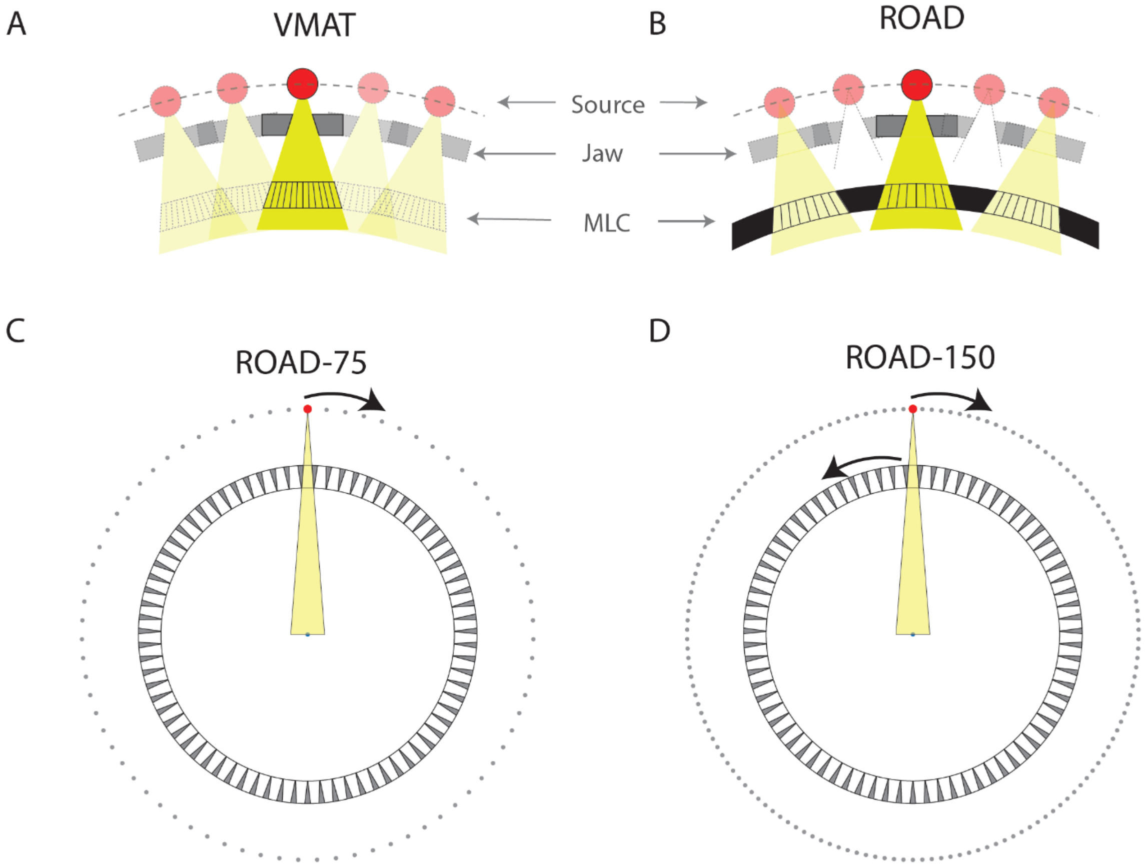


Figure 4-36 Demonstration of VMAT and ROAD.

4.6.2.2 Direct Aperture Optimization Formulation

The direct aperture optimization for FLASH therapy is formulated as

$$\begin{aligned}\underset{f,c,u}{\text{minimize}}\quad & \underbrace{\frac{1}{2}\left\|W\big((Af)-d_0\big)\right\|_2^2}_{\text{fidelity term}}+\underbrace{\left(\lambda_x\|D_xf\|_1+\lambda_y\left\|D_yf\right\|_1\right)}_{\text{anisotropic TV term on } f}\\ & +\underbrace{\frac{\gamma}{2}\left(\left\|\text{diag}(u)^{1/2}(f-c)\right\|_2^2+\left\|\text{diag}(1-u)^{1/2}f\right\|_2^2\right)}_{\text{single segment term}}\\ \text{subject to}\quad & 0\le f\le I_{\max},0\le c\le I_{\max},u\in\{0,1\}^n,u\in S_d,\end{aligned}$$

Equation 4-15

where f is the vectorized fluence map, c is the fluence intensity, and u is the vectorized aperture shape. u equals to 1 where the aperture exists, i.e., the beamlet is opened, and 0 elsewhere. The set S_d is defined to include all deliverable apertures. In the case of ROAD-150, the apertures in the set S_d are also required to satisfy the MLC speed constraint. $I_{\max}$ is the maximum intensity of the fluence map at each control point, which is limited by the maximum dose rate.

A is the fluence-to-dose transformation matrix, which converts the vectorized fluence map to the corresponding volumetric dose. W is the diagonal weighting matrix, where the diagonal elements are the weightings of the structures of interest. d_0 is the ideal dose with the prescription dose in the PTV and zero elsewhere. The least-squares fidelity term attempts to find the optimal fluence map f by penalizing the difference between the planning dose Af and the ideal dose d_0. In the second total variation (TV) term, D_x and D_y are the finite-difference matrices in the directions parallel and orthogonal to the MLC leaf motion, respectively. The anisotropic TV term encourages the piecewise continuity of fluence map f, and the amount of smoothness is controlled by λ_x and λ_y. The single segment term simplifies the final fluence map f to contain only a single segment at each control point. It pushes the fluence map f towards a uniform intensity value c within the aperture ($u = 1$, the beamlet is opened), and towards 0 outside the aperture ($u = 0$, the beamlet is closed). The single segment constraint is attained by gradually increasing the weighting parameter γ in the optimization.

4.6.2.3 Algorithm

The optimization problem in Equation 4-15 was solved using an alternating optimization approach that has been investigated in our previous studies[35,38,39,184]. In each optimization iteration, the algorithm runs through 3 modules and optimizes with respect to each of the three optimization variables f, c, and u, independently, holding the other two fixed. The process is repeated until f converges to $c \odot u$ (the elementwise/Hadamard product), indicating that single segment constraint is achieved. For ROAD-150, the optimization is initialized with the ROAD-75 plan generated using the same hyperparameters.

In the module optimizing with respect to the aperture u, the problem reduced to a graph optimization problem with a linear objective:

$$\begin{aligned} \underset{\{u_\theta\}_{\theta=0}^{n}}{\text{minimize}} \quad & \sum_{\theta} \langle q_\theta, u_\theta \rangle, \\ \text{subject to} \quad & u_\theta \in \{0,1\}^n, u \in S_d, \;\; \theta = 1,2,\dots,n_\theta \\ & (q_\theta)_j = \big((f_\theta)_j - c_\theta\big)^2 - (f_\theta)_j^2 \end{aligned}$$

Equation 4-16

where θ is the index for beam angle, and j is the index for elements in the corresponding vector. The minimization is equivalent to a simplified travelling salesman problem on a directed graph with a rectangular structure. To reduce the computation complexity, we first solved Equation 4-16 without the constraint $u \in S_d$, and then project the resultant u^* to the feasible set S_d.

In the modules optimizing with respect to the fluence map f and the intensity c, the optimization problem was solved with (FISTA)[27].

4.6.2.4 FLASH dose

When delivering a treatment plan, the physical dose changes both spatially and temporally. The spatial-dependency and time-dependency of delivered dose can be computed based on an existing treatment plan $f(t_j)$:

$$d_i(t_j) = A_i \cdot f(t_j),$$

where t_j is the time when the beam j was delivered, $f(t_j)$ is the fluence map of beam j, A_i is the dose matrix of voxel i, and $d_i(t_j)$ is the dose to voxel i at time t_j.

Similar to the physical dose, spatial and temporal variations also apply to the physical dose rate. The same physical dose could result in different FLASH effects, depending on the dose rate at each voxel. For example, if the majority of the dose delivered to a voxel was under a high dose rate, the FLASH effect within the voxel would be substantial. On the contrary, if the majority of the dose delivered to a voxel was under a low dose rate, the FLASH effect could be minimal. To evaluate the FLASH effect of the ROAD plan, the physical dose rate at each voxel needs to be evaluated at each time point.

We use the FLASH dose (FD) to denote a portion of the physical dose delivered under high dose rate, which could potentially achieve a greater FLASH effect. The FLASH dose is defined as the cumulative physical dose delivered with an average dose rate higher than a threshold

R for a minimum time interval of Δt ($R = 40$ Gy/s and $\Delta t = 100ms$ in this study). For voxel i, T_i is a set of time points with FLASH dose delivery:

$$T_i = \left\{ t_s \middle| \frac{\sum_{v=i}^{v=j} d_i(t_v)}{\sum_{v=i}^{v=j} t_v} \geq R,\ t_j - t_i \geq \Delta t,\ i \leq s \leq j \right\}.$$

The FLASH dose at voxel i is defined as

$$\mathrm{FD_i} = \sum_{s \in T_i} d_i(t_s).$$

Under the same physical dose, a higher FD indicates less biological damage to the normal tissue.

4.6.2.5 FLASH Biological Equivalent Dose

A quantitative biological model, the Radiolytic Oxygen Depletion (ROD) model (Pratx and Kapp, 2019), is adapted to evaluate the biological effect of the ROAD model. We assumed the same parameters as Pratx et al.[185] in this study.

The radioprotective effect of FLASH irradiation has been connected with decreased radiosensitivity of normal tissue cells due to transient ROD [171,173,185–190]. The effect of oxygen on radiosensitivity can be quantified according to the Oxygen Enhancement Ratio (OER), defined as the ratio of the dose in anoxia to the dose under a certain oxygen tension p to achieve the same biological effect. OER can be parameterized as:

$$\mathrm{OER}(p) = 1 + \alpha\left(1 - e^{-\psi \cdot p}\right)$$

Equation 4-17

where $\alpha = 1.63$, $\psi = 0.26\ \mathrm{mmHg}^{-1}$, and p is the transient oxygen tension.

To compute the transient oxygen tension in ROAD plan delivery, we divided the delivery process into two parts: within-pulse and between-pulse, and discuss them separately.

Within the high-intensity short-pulse (> 10^3 Gy/s), the effect of oxygen metabolism and tissue diffusion is negligible, and the transient oxygen tension has a linear relationship on the dose input [187,191]:

$$\boldsymbol{p_i^+} = \boldsymbol{p_i^-} - \mathbf{L} \cdot \boldsymbol{d_i}$$

Equation 4-18

where i is the pulse index, d_i is the dose of pulse i, L is the ROD rate (0.42 mmHg Gy^{-1}) [191], p_i^- is the transient oxygen tension right before pulse i, and p_i^+ is the transient oxygen tension immediately after pulse i.

The oxygen tension ($p_i^+ \rightarrow p_{i+1}^-$) between the two dose-pulses (pulse $i \rightarrow$ pulse $i+1$) changes with oxygen diffusion and tissue metabolisms. Following the computational model proposed by Pratx et al. [190], we assumed an infinitely long capillary (radius $r_0 = 3\mu m$) and a constant oxygen tension within the capillary ($p_{cap} = 40mmHg$). The effects of different capillary oxygen levels have been discussed in Figure 6 in Pratx et al. [190]. With polar symmetry, the oxygen diffusion and tissue metabolism is modeled in polar coordinates

$$\frac{\boldsymbol{\partial p}}{\boldsymbol{\partial t}} = \mathbf{a}\frac{\mathbf{1}}{\boldsymbol{r}}\frac{\boldsymbol{\partial}}{\boldsymbol{\partial r}}\left(\boldsymbol{r}\frac{\boldsymbol{\partial p}}{\boldsymbol{\partial r}}\right) - \mathbf{m},$$

Equation 4-19

where the diffusion constant $\mathrm{a} = 2 \cdot 10^{-5}\mathrm{cm}^2\mathrm{s}^{-1}$ [192], oxygen metabolism rate $\mathrm{m} = 3\mathrm{mmHg\ s}^{-1}$, and r is the distance of the cell from the capillary. The boundary condition is set by enforcing $p(r \le r_0) = 40\mathrm{mmHg}$. Given the transient oxygen tension $p_i^+(r)$ right after pulse i, the transient oxygen tension $p_{i+1}^-(r)$ right before pulse $i+1$ can be computed iteratively using a finite difference method. The initial state $p_0(r)$ is set as the steady state of Equation 4-19, which can be solved by setting $\frac{\partial p}{\partial t} = 0$.

The transient oxygen tension $p(t, r)$ in ROAD plan delivery can be computed voxelwise by iteratively applying Equation 4-18 ($p_i^- \rightarrow p_i^+$) and solving Equation 4-19 ($p_i^+ \rightarrow p_{i+1}^-$) using the sequential dose pulses received by each voxel.

In conventional VMAT plan delivery, the oxygen tension stays at the initial condition $p_0(r)$. We define the transient radiosensitivity S as the ratio of the transient dose in ROAD to the dose under conventional dose rate to achieve the same biological effect:

$$S(t) = \frac{\mathrm{OER}\big(p(t)\big)}{\mathrm{OER}(p_0)}.$$

Equation 4-20

For simplicity, we assumed 75μm from the capillary ($r = 75\mu\mathrm{m}$) for radiosensitivity evaluation. Figure 4-37 shows examples of the time series plots of the dose pulses and the transient radiosensitivity S, assuming different pulse sequences (pulse sequence 1 to 4) and the same total dose 25Gy. Note that varying delivery time is assumed in this figure (from submillisecond to 15s) to illustrate the different biological effects resulted from different delivery schemes. With an incident dose, transient oxygen tension drops, OER decreases, and subsequently, the transient radiosensitivity S falls off. Conversely, without irradiation, the radiosensitivity S gradually rebounds until it reaches 1. Under different dose patterns (pulse sequences 1 to 4), the transient oxygen tension $p(t)$ could be different, leading to different transient radiosensitivity S, and different overall biological effects, even if the total dose is the same.

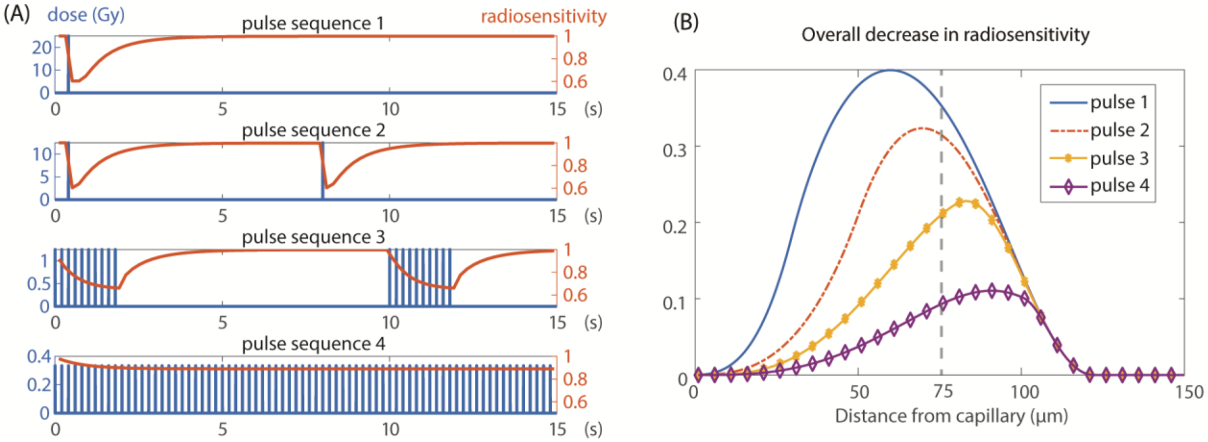


Figure 4-37 (A) Time series plot of transient radiosensitivity S under different dose inputs (pulse sequences 1 to 4), assuming a total dose of 25Gy. (B) Overall decrease in radiosensitivity as a function of the distance from capillary, plotted for all four pulse sequences. $r = 75\mu m$ (indicated by the dashed line) were assumed for computing the transient radiosensitivity S in (A).

The fractional cell kills from an infinitesimal dose dD with a transient radiosensitivity S is

$$\frac{dN}{N} = \beta S\, dD,$$

Equation 4-21

where N is the number of surviving cells, and β is a constant. The FLASH biological equivalent dose (FBED) is defined as the dose to achieve the same biological effect in conventional dose rate irradiation, which is equivalent to the integral of the radiosensitivity over dose:

$$\mathrm{FBED} = \int S dD = \frac{\int \mathrm{OER}(p(D))dD}{\mathrm{OER}(p_0)}.$$

Equation 4-22

In general, the FLASH effect becomes more evident with increasing dose, due to an increased amount of oxygen depletion. In addition, the effect is also dependent on the dose sequence as illustrated in Figure 4-37. This computational model evaluates the FBED individually for each voxel, based on the incident dose at each voxel level, as a comprehensive evaluation including both the total delivered dose and the dose sequences.

Note that the FBED calculation only applies to normal cells. We assumed the FBED of the PTV is equivalent to its physical dose for simplicity.

4.6.2.6 Evaluation

The single-arc ROAD plans were compared with clinical two-arc VMAT plans on four representative patients, including a glioblastoma multiforme case (GBM), a lung cancer case (LNG), a prostate cancer case (PRT), and a Head and Neck cancer case (H&N). The VMAT plans were planned on the Eclipse treatment planning system using two coplanar arcs with 90 degrees collimator rotation. The dose matrix for ROAD was obtained using a convolution/superposition dose calculation algorithm with 10 MV X-ray polyenergetic kernels [100,134]. The gantry angles were sampled with 4.8 degree and 2.4 degree spacing for ROAD-75 and ROAD-150, respectively. For VMAT, the gantry angle was sampled with 2 degree spacing (180 control points in total). The beamlet size for dose calculation was 0.5×0.5 cm^2 for all plans. The beam field of view (FOV) was 15×20 cm^2 for ROAD, and 20×20 cm^2 for VMAT. The dose voxel size was 0.25×0.25×0.25 cm^3.

The dose rate was calculated by converting the machine output in Monitor Units (MU) to dose and dividing by the beam time:

$$\text{dose rate} = \frac{\text{Output (MU)}}{\text{time (s)}} * \frac{1\ (\text{Gy})}{100\ (\text{MU})}.$$

Conversion from fluence intensity to MU was calibrated by calculating unit-fluence dose in water using the same dose calculation algorithm and measuring the dose at dmax.

For ROAD plans, the physical dose, the FLASH dose, and the FBED were evaluated. The FLASH dose and FBED computation assumed 25 Gy prescription dose to the PTV in a single fraction. The physical dose of ROAD plans was compared with that of the corresponding clinical VMAT plans. The FLASH dose was computed for all normal-tissue voxels. The distribution of FLASH dose vs. physical dose was evaluated to show the potential FLASH effect of the ROAD plans. The FBED were computed following the biological model, assuming nominal parameter values and 25% uncertainty ($p_{cap} = 40 \pm 10\ mmHg$, $L = 0.42 \pm 0.105\ \mathrm{mmHg\ Gy^{-1}}$, and $\mathrm{a} = 2 \cdot 10^{-5} \pm 5 \cdot 10^{-6} \mathrm{cm^2 s^{-1}}$).

PTV statistics including PTV D98, D2, and PTV homogeneity (defined as $\frac{\mathrm{D95}}{\mathrm{D5}}$) were evaluated. R50, defined as the 50% isodose volume divided by the target volume, was also assessed to quantify the dose compactness. The integral dose was evaluated as total dose spillage. For OARs, the maximum dose (Dmax) and the mean dose (Dmean) were obtained. The maximum dose is defined as D2 (the dose at 2% of the structure volume), following the ICRU-83 report [101].

4.6.3 Results

4.6.3.1 Plan information

Case	Method	Prescription dose (Gy)	PTV volume (cc)	Single-fraction dose (Gy)	Estimated delivery time (s)	Per-beam dose rate (Gy/s)	
						Mean	Max
GBM	VMAT	25	6.359	2	132		
	ROAD-75			25	1	94.66	150
	ROAD-150			25	1	141.25	300
LNG	VMAT	50	137.7	2	132		
	ROAD-75			25	1	61.34	150
	ROAD-150			25	1	89.27	300
PRT	VMAT	40	84.17	2	132		
	ROAD-75			25	1	71.64	150
	ROAD-150			25	1	111.84	300
H&N	VMAT	70 (56)	132.9 (90.7)	2	132		
	ROAD-75			25	1	77.28	150
	ROAD-150			25	1	105.13	300

Table 4-21 Prescription doses, PTV volumes, single-fraction dose, estimated delivery time per fraction, and mean and maximal per-beam dose rate for all plans.

Error! Reference source not found. reports the prescription doses, PTV volumes, single-fraction dose, estimated delivery time, and mean and maximal per-beam dose rate for all ROAD plans. The single fraction dose is assumed 2 Gy for VMAT and 25 Gy for the ROAD plans. For the H&N patient, the single fraction dose of the ROAD plan is 25Gy for the partial PTV with the highest prescription dose (PTV70Gy) and scaled accordingly for the other PTV (PTV56Gy). On average, the mean dose rate of ROAD-75 and ROAD-150 are 76.2 Gy/s and 112 Gy/s, respectively, and the maximal dose rate are 150Gy/s and 300 Gy/s. The maximal dose rate is limited by the optimization constraint that enforces a maximum of 2Gy dose per control point. Note that the provided dose rate is amortized over the beam time, while the instantaneous dose rate (within-pulse dose rate) should be several orders higher, depending on the pulse duration. In addition, the instantaneous/maximal/mean dose rates are different from the dose rate distribution within the patient body, while the latter was used for voxelwise quantification of the FLASH dose and the FBED.

4.6.3.2 Physical dose

Figure 4-38 presents the isodose comparison of the physical dose of VMAT, ROAD-75, and ROAD-150 for the H&N and the PRT patient. ROAD-75 and ROAD-150 achieved similar OAR dose, while both significantly reduced OAR dose compared with VMAT. ROAD substantially reduced physical dose to the spinal cord and mandible for the H&N patient, and to the rectum, seminal vesicle, and penile bulb for the PRT patient.

Figure 4-39 compares the physical dose DVHs of VMAT (solid), ROAD-75 (dotted), and ROAD-150 (dashed) for all patients. PTV D95 is normalized to prescription dose in all plans. ROAD-150 improved PTV homogeneity compared with the clinical VMAT plans and reduced physical dose to OARs.

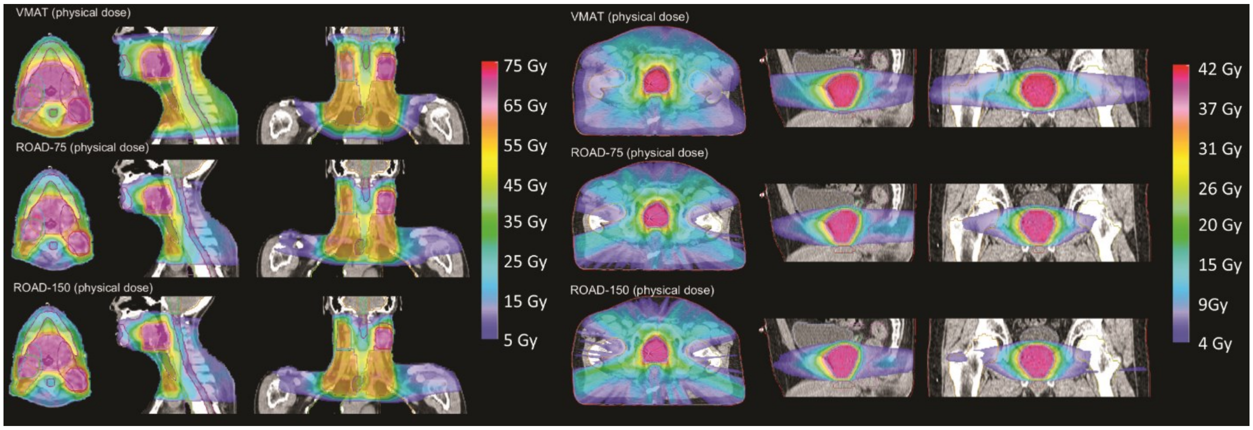


Figure 4-38 Isodose comparison of VMAT physical dose, and physical dose of ROAD-75 and ROAD-150 for the H&N and the PRT patient.

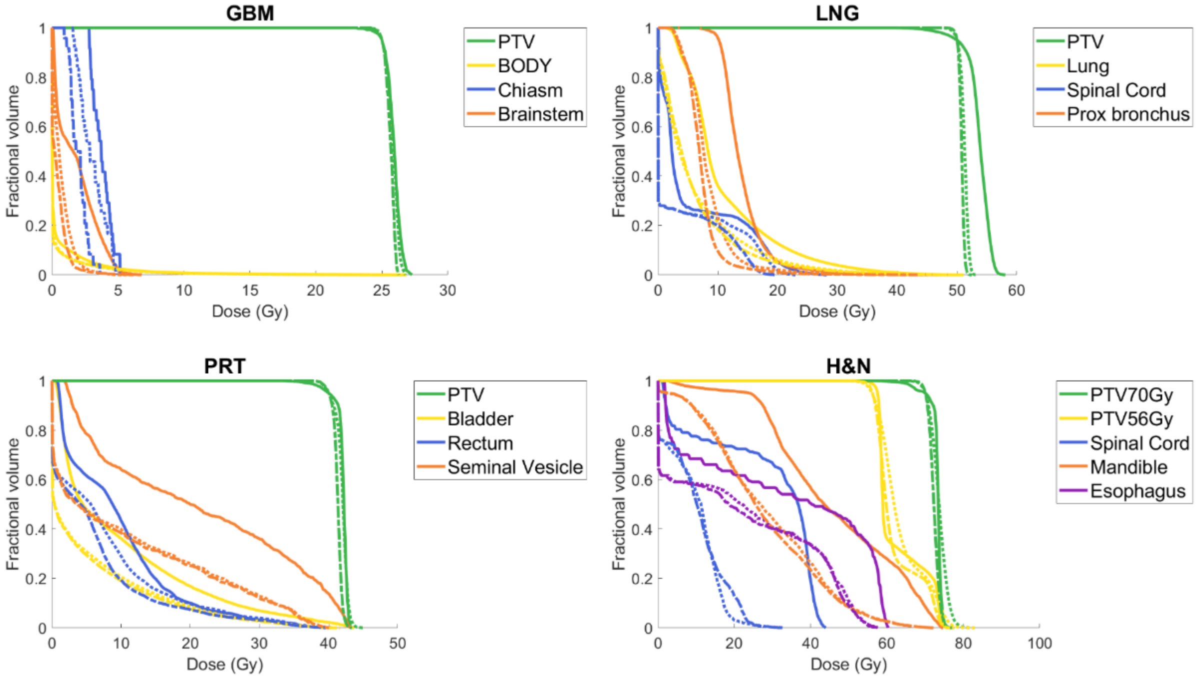


Figure 4-39 DVH comparison of VMAT physical dose (solid), ROAD-75 physical dose (dotted), and ROAD-150 physical dose (dashed) for all plans with selected OARs. PTV D95 is normalized to prescription dose in all plans.

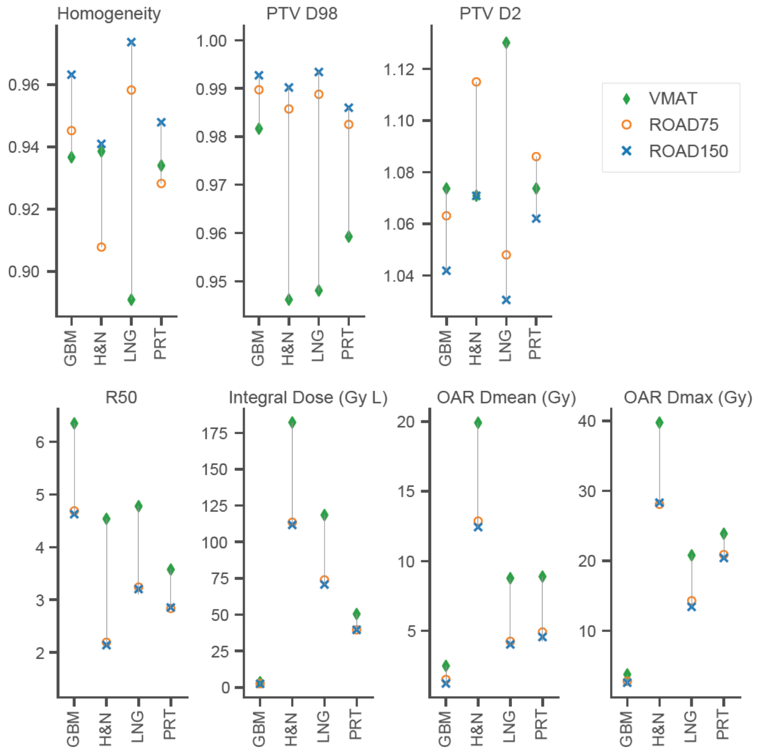


Figure 4-40 Comparison of VMAT (green diamonds), ROAD-75 (orange circles), and ROAD-150 (blue crosses) on PTV statistics (Homogeneity, PTV D98 and D2 normalized by the prescription dose), R50, Integral Dose normalized by the prescription dose, and OAR maximum and mean doses.

Figure 4-40 reports the PTV and OAR statistics of the physical dose, comparing VMAT (green diamonds), ROAD-75 (orange circles), and ROAD-150 (blue crosses). Compared with clinic VMAT plans, ROAD-150 substantially improved PTV homogeneity, increased PTV minimum dose (D98), and reduced PTV maximum dose (D2). The PTV homogeneity improved by 3%, PTV D98 increased by 3%, and PTV D2 decreased by 3%. The R50, integral dose, and OAR mean and maximal dose were markedly reduced in the ROAD plans. Compared with clinical VMAT, ROAD-150 reduced the [max, mean] OAR physical dose by [4.8Gy, 6.3Gy]. The average R50 and integral dose of [VMAT, ROAD-75, ROAD-150] are [4.8, 3.2, 3.2] and [89, 57, 56] Gy×Liter, respectively.

4.6.3.3 FLASH dose

Figure 4-41 shows the heat map of physical dose and FLASH dose distribution of normal tissue voxels receiving physical dose greater than 10Gy, based on the ROAD-150 plans for all patients, assuming 25 Gy prescription dose in a single fraction. The heat map was normalized columnwise. Overall, the ratio of FLASH dose to the physical dose is around 50% to 70%, indicating that the majority of the delivered physical dose could trigger the FLASH effect.

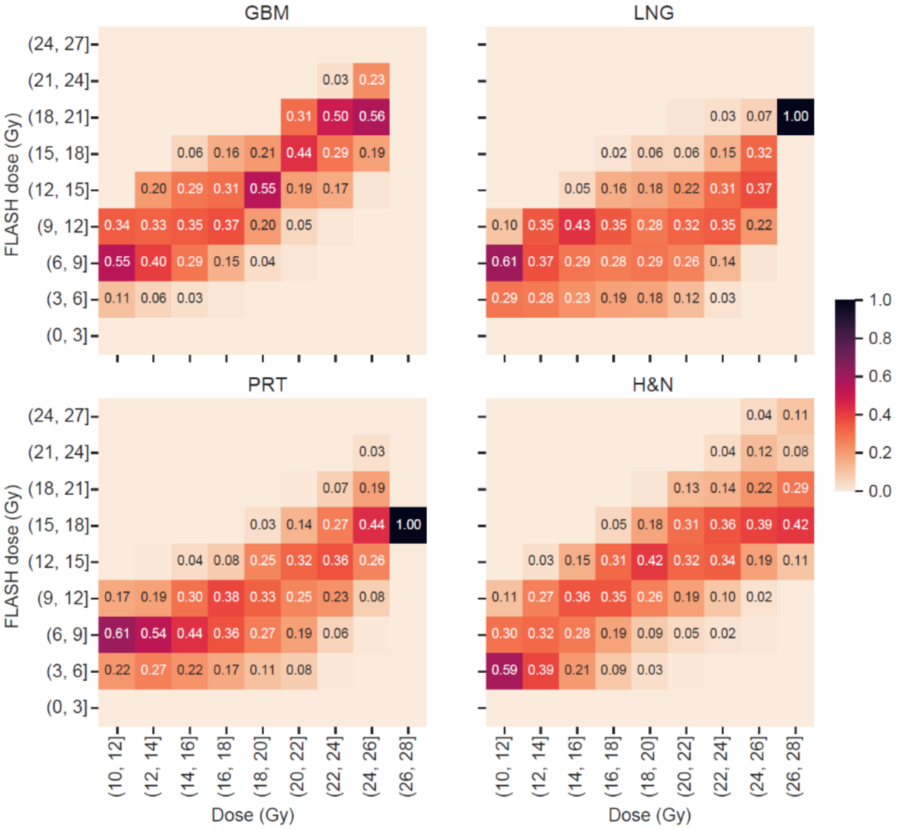


Figure 4-41 Heat map of dose and FLASH dose distribution of normal tissue voxels receiving physical dose greater than 10Gy, based on the ROAD-150 plans for all patients, assuming 25 Gy prescription dose in a single fraction. The heat map was normalized columnwise.

Figure 4-42 compares the physical Dose Volume Histogram (solid) and the FLASH Dose Volume Histogram (dotted) of the ROAD-150 plan for the H&N patient. The FLASH dose takes up a large portion of the physical dose, especially in the high-dose region. For example, 30% of the trachea received no less than around 18 Gy of physical dose within a single fraction. In the meantime, 30% of the trachea received no less than around 13 Gy FLASH dose, indicating the biological damage to the trachea could be substantially lower with the ROAD delivery.

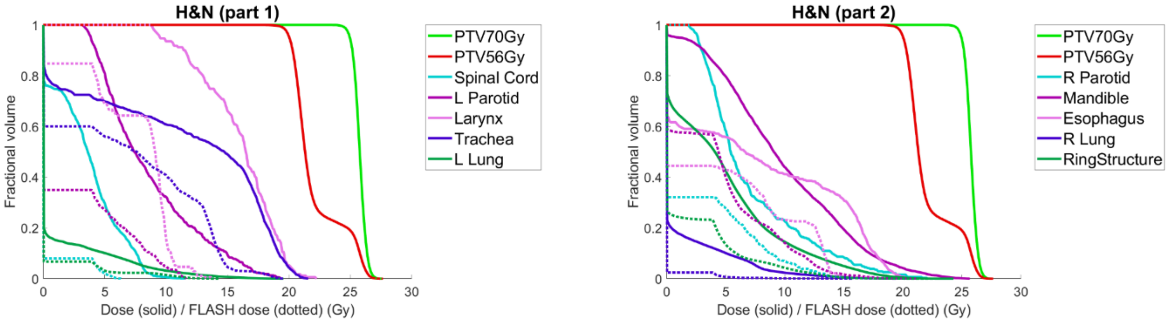


Figure 4-42 Dose Volume Histogram (solid) and FLASH Dose Volume Histogram (dotted) of the ROAD-150 plan for the H&N patient.

4.6.3.4 FLASH Biological Equivalent Dose

Figure 4-43 reports the statistical comparison of the physical dose (orange circles) and the FBED (purple squares) of ROAD-75 plans for all patients, assuming nominal parameter values and 25% uncertainty. The FBED under nominal parameter values substantially reduced R50, integral dose, and maximum and mean OAR doses compared with the physical dose, indicating potentially significant biological gains with ROAD plans. However, the large error bars also suggest large uncertainties due to the uncertainties in the biological model. Note that the FLASH effects are most evident in the high-dose regions, where the normal tissues receive a substantial dose that depletes transient oxygen and triggers the FLASH effect. For example, the H&N case has a significantly higher OAR mean and maximal dose than other cases, and subsequently, it shows the most significant FLASH effect in OAR mean and maximal dose. However, all four cases showed substantial FLASH effects, as indicated by the sharp reduction of R50.

Figure 4-44 compares the physical Dose Volume Histogram (solid) and the FLASH Biological Equivalent Dose Volume Histogram (dotted) of the ROAD-150 plan for the H&N patient. The FBED were computed assuming nominal parameters and a single-fraction dose of 25Gy. The

OAR FBED was markedly reduced compared with the corresponding physical dose, indicating the FLASH effect could result in less OAR damage in the ROAD plans.

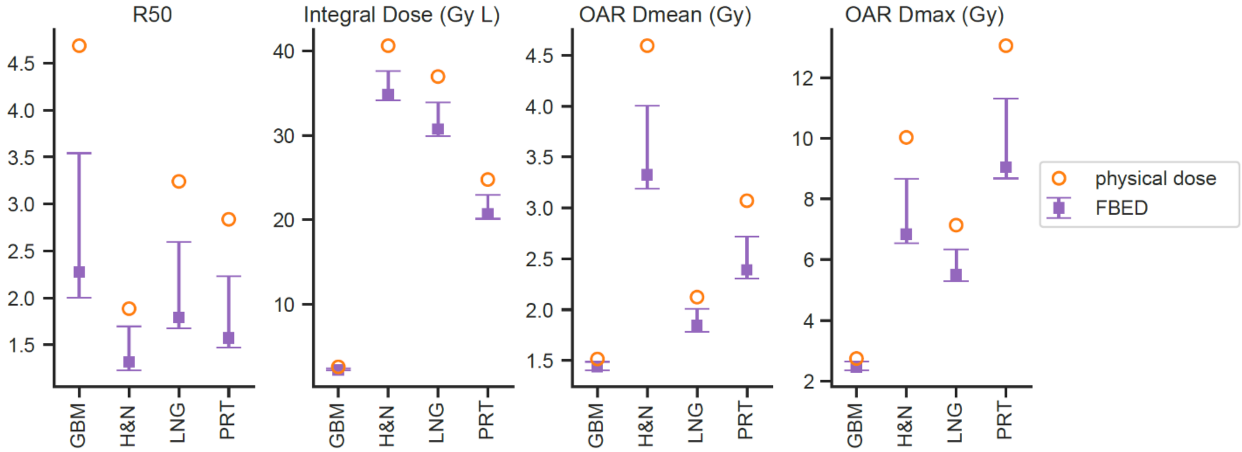


Figure 4-43 Comparison of physical dose (orange circles) and the FLASH Biological Equivalent Dose (purple squares) with error bar of ROAD-75 plans for all patients on R50, Integral Dose normalized by the prescription dose, and OAR maximum and mean doses.

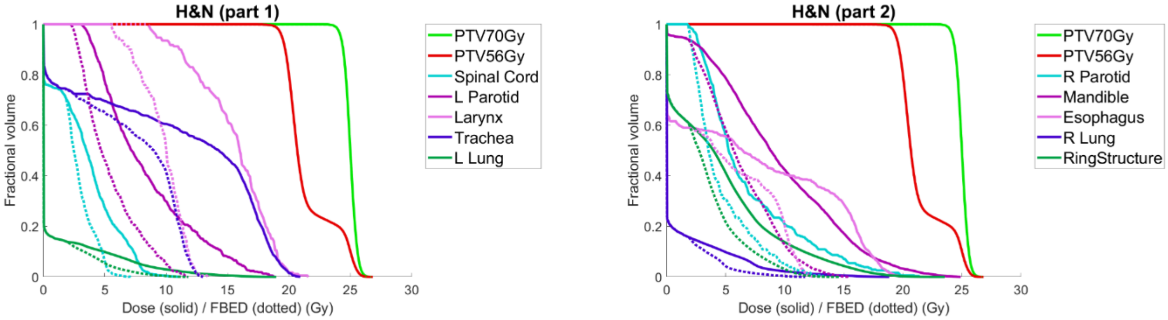


Figure 4-44 Dose Volume Histogram (solid) and FLASH Biological Equivalent Dose Volume Histogram (dotted) of the ROAD-150 plan for the H&N patient, assuming nominal parameters in the FBED model and a single-fraction dose of 25Gy.

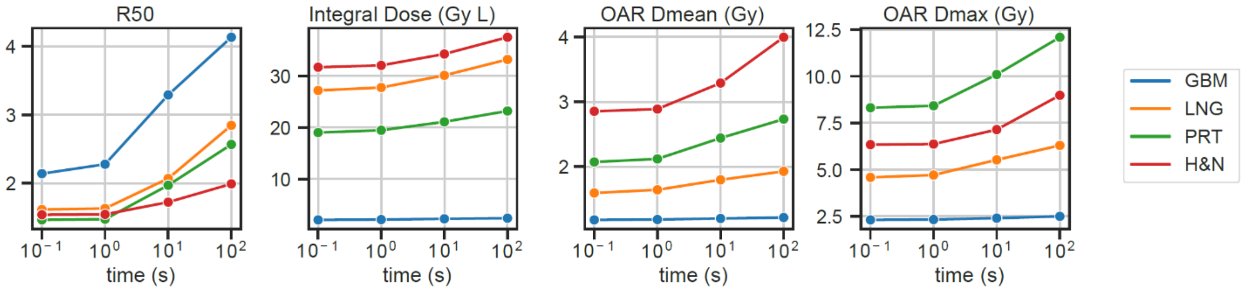


Figure 4-45 Time dependence of the ROAD-150-FBED, evaluated for all patients over four metrics, including R50, Integral Dose normalized by the prescription dose, and OAR maximum and mean doses, assuming nominal parameters in the FBED model and a single-fraction dose of 25Gy. The delivery time ranges from 0.1s to 100s. The OAR maximum and mean doses were the averaged values across all OARs for each individual patient.

Figure 4-45 shows the time dependence of the ROAD-150-FBED. As delivery time increases from 0.1s to 100s, the R50, integral dose, and OAR Dmean and Dmax increase, reaching that of the physical dose with 100s delivery time. The dose compactness and OAR sparing substantially improved with shortened delivery time, and plateaued with a delivery time less than 1s. Note that the results are dependent on the parameters in the FBED model. For example, if a greater diffusion constant were assumed, the oxygen would be replenished at a faster rate, leading to a shorter time threshold of observing the FLASH effect.

4.6.4 Discussion

Existing treatment machines with a single MLC module are mechanically too slow for FLASH radiotherapy. This study proposes the novel ROAD design, which mounts 75 MLC modules on a separate ring and pre-shapes the apertures to achieve ultrafast dose modulation. We can find an analogy in computer science, where a *space-time* tradeoff manifests through solving a problem in less time by using more storage space or in limited space by spending extra time. ROAD and VMAT both enforce a single fluence map segment per beam, but they differ in delivery speed, the number of control points, and aperture continuity constraints: VMAT requires similar aperture shapes for adjacent beams, and ROAD-150 has the leaf speed constraint on the beams that share the same MLC modules. Without aperture continuity constraint on adjacent beams, ROAD plans can be delivered within 1s, providing a viable solution for clinical FLASH radiotherapy.

Compared with clinical VMAT plans, ROAD-150 improved PTV homogeneity and substantially reduced OAR dose. The improvement in physical dose is attributed to the integrated optimization framework, which has been applied to VMAT optimization and was shown to outperform the progressive sampling optimization method used in clinical VMAT [35]. The potential FLASH effect further increases the therapeutic window in addition to the physical dose improvement. Apart from the improved physical dose, ROAD maintained the per-beam dose under 2Gy for practical linac design and achieved a markedly higher mean dose rate. Previously reported studies indicated that the FLASH effect was observed when the mean dose rate is greater than 40 Gy/s. In this study, assuming a single-fraction dose of 25Gy and treatment time of 1s, the average per-beam dose rate is 76.2 Gy/s and 112 Gy/s for ROAD-75 and ROAD-150, respectively. 50% to 70% of the physical dose was delivered with equal to or greater than the FLASH threshold dose rate for ROAD to take advantage of the additional FLASH biological effects in addition to the improved physical dose. We note that the estimated percentage could vary with our continuously improving knowledge of the FLASH dose rate.

The ROAD system is a conceptual software and hardware platform for ultrafast radiotherapy delivery, where the optimization is based on the physical dose. Besides the main point, to demonstrate the feasibility of incorporating a quantitative FLASH biological model into planning, we selected the ROD model (Pratx and Kapp, 2019) as an example to evaluate the biological effectiveness under various operational conditions. To employ ROD, we made several extensions to the original model. First, the original ROD paper (Pratx and Kapp, 2019) only discusses the FLASH effects on a small subvolume, while we apply the model to the whole patient body. Solving the large-scale differential equation for a CT image containing tens of millions of voxels using parallel computing is not trivial. Second, the ROD paper focuses on the decreased radiosensitivity with varying parameters, whereas the current study calculates FBED based on the dose-averaged radiosensitivity. Third, we tailored the biological model to fit the ROAD model.

The FD and FBED models attempt to address one common challenge in FLASH radiotherapy, which is to evaluate the FLASH effect for radiation with complex spatial and temporal structures. In the ROAD plan delivery, the pulse dose rate is different from the instantaneous dose rate distributed within the patient body. For proton FLASH in the shoot-through mode, although a single proton field can be delivered with a high dose rate, the time to deliver multiple fields for conformal dose distribution will result in a low average dose rate, which would be significantly lower than the proposed FLASH system. The quantitative FD and FBED model is one way to reconcile the apparent conflict between the instantaneous dose rate seen by a subvolume and the average dose rate.

Nevertheless, we note that the radiolytic biological model in FBED computation only partially explains the FLASH effect [193], and its correctness is pending further experimental validation. For example, a limitation of the ROD model is that the model assumes the cells sufficiently far from capillaries (thus experiencing hypoxia) to experience significant FLASH effects. The limitation indicates an incomplete understanding of the FLASH mechanism. There are also large uncertainties in the parameters, such as the diffusion constant and the ROD rate, and a lack of threshold on the instantaneous dose rate to trigger the FLASH effect. These uncertainties could lead to large uncertainties in estimating the FBED. Emerging evidence may result in revised or new quantitative FLASH biological models. We will update the evaluation of ROAD accordingly.

Different from PHASER [183], ROAD is designed based on scaling up existing technologies. For example, existing MLCs are used with a larger quantity. The high-speed slip ring gantry has been demonstrated by Reflexion (Hayward, CA) [194]. Nonetheless, there are challenges beyond the scope of the concept paper. The significantly larger high output linac will be more challenging to integrate with the high-speed gantry. It would not be trivial to design an X-ray target system that can withstand the high beam power. Moreover, to maximize the output, the linac will be flattening-filter-free (FFF). As a VMAT system, the dose rate inhomogeneity will be managed in ROAD as linear constant multipliers in the inverse optimization. Because FFF-VMAT has been extensively studied with the minimal dosimetric difference compared with flattened-field VMAT, the physical dose impact is expected to be minimal. On the other hand, since the linac output is defined by the peak output at the center of the field, the effective dose rate will be lower, particularly for larger targets. We would evaluate the change in FBED with a FLASH biological model or increase the linac output to compensate for the reduced dose rate. Finally, the RF power system is to be optimized for the burst-mode operation. We will develop individual engineering solutions for overcoming these challenges in future work.

4.6.5 Summary

The novel ROAD design achieves ultrafast dose delivery and improves physical dosimetry compared with clinical VMAT, providing a potentially viable engineering solution for X-ray FLASH radiotherapy.

5 APPENDIX

5.1 The FISTA Algorithm

The FISTA algorithm solves problems in the following canonical form

$$minimize\ F(x) + G(x),$$

Equation 5-1

where F is a differentiable convex function with a Lipschitz continuous gradient, and G is a convex function which has a proximal operator that can be evaluated efficiently. The proximal operator [26] of a function G with step size t is defined by

$$prox_{tG}(x) = \underset{z}{argmin}\left(G(z) + \frac{1}{2t}\|z - x\|_2^2\right).$$

Equation 5-2

The pseudocode for the FISTA with line search algorithm is summarized in Table 5-1, where evaluation of the gradient of F and the proximal operator of G are required at each iteration, and the function value of F is also assessed to obtain the optimal step size through a line search method.

Pseudocode for FISTA with line search

```
Initialize $x_0 := 0, v_0 := x_0, t_0 > 0, r_1 > 1, r_2 > 1$
for $k = 1, 2, \dots$ do
        $t \coloneqq r_1 t_{k-1}$
        Repeat
            $\theta \coloneqq \begin{cases} 1 & if \quad k = 1 \\ positive\ root\ of\ t_{k-1}\theta^2 = t\theta_{k-1}^2(1-\theta) & if \quad k > 1 \end{cases}$
            $\boldsymbol{y \coloneqq (1-\theta)x_{k-1} + \theta v_{k-1}}$
            $x \coloneqq prox_{tG}\left(y - t\nabla F(y)\right)$
          break if $\boldsymbol{F(x) \leq F(y) + < \nabla F(y), x - y > + \frac{1}{2t}\|x - y\|_2^2}$
        $t \coloneqq t/r_2$
    $\boldsymbol{t_k \coloneqq t}$
    $\theta_k \coloneqq \theta$
    $\boldsymbol{v_k \coloneqq x_k + \frac{1}{\theta_k}(x - x_k)}$
    break if $\frac{\|x - x_k\|}{\|x_k\|} \leq \epsilon$
    $\boldsymbol{x_k \coloneqq x}$
end for
return $x$
```

Table 5-1 FISTA with line search

5.2 Prox-operator calculations

5.2.1 Group sparsity term

Evaluation of the proximal operator of $\Theta(x) = \rho\|x\|_2 + I_+(x)$ at $\hat{x}$ is equivalent to finding the minimizer for the optimization problem

$$\underset{x}{minimize} \quad \|x\|_2 + \frac{1}{2\rho t}\|x - \hat{x}\|_2^2$$

$$subject\ \ to \qquad x \geq 0.$$

Equation 5-3

For $\hat{x}_i \leq 0$, the corresponding minimizer always takes its ith element as $x_i = 0$. Thus this problem is equivalent to the following unconstrained optimization problem

$$\underset{x}{minimize} \ \|x\|_2 + \frac{1}{2\rho t}\|x - max(\hat{x}, 0)\|_{2.}^2$$

Equation 5-4

This is no different from evaluation of the prox-operator of the l_2 norm. Thus

$$prox_{t\Theta}(\hat{x}) = prox_{\rho t\|\cdot\|_2}\big(max(\hat{x},0)\big) = \hat{x} - P_{\rho t}\big(max(\hat{x},0)\big),$$

Equation 5-5

where $P_{\rho t}(x)$ projects x onto a Euclidean ball with radius ρt.

5.2.2 $|x|^{\alpha}$

The proximal operator of $|x|^{\alpha}$ is

$$\left(\boldsymbol{Prox}_{w|x|^{\alpha}}(\boldsymbol{z})\right)_{\boldsymbol{i}} = \underset{\boldsymbol{x}}{\boldsymbol{argmin}}\left(\boldsymbol{\eta}|\boldsymbol{x}|^{\boldsymbol{\alpha}} + \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{w}}(\boldsymbol{x}-\boldsymbol{z}_{\boldsymbol{i}})^{\mathbf{2}}\right)$$

Equation 5-6

Exact analytic solutions to exist for $\alpha = 0, \frac{1}{2}, \frac{2}{3}, 1$. For other values of α, the proximal operators could be evaluated numerically with an iterative approach such as the Newton's method [88].

For $\alpha = 0$, the proximal operator reduces to a hard thresholding operation.

$$\left(Prox_{w|x|^{\alpha}}(z)\right)_i = \begin{cases} z_i, & if\ |z_i| \geq \sqrt{2\eta w} \\ 0, & if\ |z_i| < \sqrt{2\eta w} \end{cases}.$$

For $\alpha = \frac{1}{2}$, the closed form solution was proposed by McKelvey [16]

$$\left(Prox_{w|x|^{\alpha}}(z)\right)_i = \begin{cases} \frac{4}{3}sin^2\left(\frac{1}{3}arccos\left(\frac{3\sqrt{3}}{4}u\right)+\frac{\pi}{2}\right), & if\ \ u \leq \frac{2\sqrt{6}}{9} \\ 0, & if\ \ u > \frac{2\sqrt{6}}{9}, \end{cases}$$

where $u = \eta w |\mathrm{z_i}|^{-\frac{3}{2}}$.

For $\alpha = \frac{2}{3}$, the solution of the proximal operator is a root of the quartic polynomial

$$\boldsymbol{x^4 - 3z_i x^3 + 3z_i^2 x^2 - z_i^3 x + \frac{8\eta^3 w^3}{27} = 0},$$

for which the analytic solution exists. The details on the proximal operator evaluation for $\alpha = \frac{2}{3}$ are shown in Table 5-2.

For $\alpha = 1$, the proximal operator results in a soft thresholding

$$\left(\mathrm{Prox}_{w|x|^{\alpha}}(\mathbf{z})\right)_i = \begin{cases} \mathrm{z_i} - \eta, & if\ \mathrm{z_i} \geq \eta, \\ 0, & if\ |\mathrm{z_i}| < \eta, \\ \mathrm{z_i} + \eta, & if\ \mathrm{z_i} \leq -\eta. \end{cases}$$

Proximal operator evaluation of $|x|^{\alpha}$ $(\alpha = 2/3)$

Compute intermediary terms $\boldsymbol{a_1, a_2, a_3, a_4}$:

$$\boldsymbol{a_1} \coloneqq \frac{\mathbf{8}\boldsymbol{w^3\eta^3}}{\mathbf{27}\hat{\mathbf{z}}^{\mathbf{4}}_{\mathbf{3}\boldsymbol{i}}},\ \boldsymbol{a_2} \coloneqq \sqrt[3]{\frac{\boldsymbol{a_1}}{\mathbf{16}} + \sqrt{-\frac{\boldsymbol{a_1^3}}{\mathbf{27}} + \frac{\boldsymbol{a_1^2}}{\mathbf{256}}}},\ \boldsymbol{a_3} \coloneqq \frac{\mathbf{5}}{\mathbf{8}} + \mathbf{2}\boldsymbol{a_2} + \frac{\mathbf{2}\boldsymbol{a_1}}{\mathbf{3}\boldsymbol{a_2}},\ \boldsymbol{a_4} \coloneqq \sqrt{\boldsymbol{a_3} - \frac{\mathbf{3}}{\mathbf{8}}}$$

Compute 4 roots $\boldsymbol{r_1, r_2, r_3, r_4}$:

$$\boldsymbol{r_1} \coloneqq \frac{\mathbf{3}}{\mathbf{4}} + \left(\boldsymbol{a_4} + \sqrt{\frac{\mathbf{9}}{\mathbf{8}} - \boldsymbol{a_3} - \frac{\mathbf{1}}{\mathbf{4}\boldsymbol{a_4}}}\right)/\mathbf{2},\quad \boldsymbol{r_2} \coloneqq \frac{\mathbf{3}}{\mathbf{4}} + \left(\boldsymbol{a_4} - \sqrt{\frac{\mathbf{9}}{\mathbf{8}} - \boldsymbol{a_3} - \frac{\mathbf{1}}{\mathbf{4}\boldsymbol{a_4}}}\right)/\mathbf{2}$$

$$\boldsymbol{r_3} \coloneqq \frac{\mathbf{3}}{\mathbf{4}} + \left(-\boldsymbol{a_4} + \sqrt{\frac{\mathbf{9}}{\mathbf{8}} - \boldsymbol{a_3} + \frac{\mathbf{1}}{\mathbf{4}\boldsymbol{a_4}}}\right)/\mathbf{2},\quad \boldsymbol{r_4} \coloneqq \frac{\mathbf{3}}{\mathbf{4}} + \left(-\boldsymbol{a_4} - \sqrt{\frac{\mathbf{9}}{\mathbf{8}} - \boldsymbol{a_3} + \frac{\mathbf{1}}{\mathbf{4}\boldsymbol{a_4}}}\right)/\mathbf{2}$$

Pick the root $\boldsymbol{r}$ as the maximum of $\{\mathbf{0}, \boldsymbol{r_1, r_2, r_3, r_4}\}$ that satisfies
$\boldsymbol{r}$ is a real number

$$\mathbf{0.5} < \boldsymbol{r} < \mathbf{1}$$

Output: $\left(\mathbf{Prox}_{\boldsymbol{w}|\boldsymbol{x}|^{\alpha}}(\mathbf{z})\right)_{\boldsymbol{i}} = \boldsymbol{rz_i}$

Table 5-2 Proximal operator evaluation of $|x|^{\alpha}$ $(\alpha = 2/3)$

6 REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;0(0). doi:doi:10.3322/caac.21492

2. Page BR, Hudson AD, Brown DW, et al. Cobalt, linac, or other: What is the best solution for radiation therapy in developing countries? *Int J Radiat Oncol Biol Phys*. 2014. doi:10.1016/j.ijrobp.2013.12.022

3. Smith FW, Hutchison JMS, Mallard JR, et al. Oesophageal carcinoma demonstrated by whole-body nuclear magnetic resonance imaging. *Br Med J (Clin Res Ed)*. 1981. doi:10.1136/bmj.282.6263.510

4. Hawkes R, Holland G, Moore WS, Worthington B. Nuclear Magnetic Resonance (NMR) Tomography of the Brain: A Preliminary Clinical Assessment with Demonstration of Pathology. *J Comput Assist Tomogr*. 1980. doi:10.1097/00004728-198010000-00001

5. Moore EA. Magnetic resonance imaging. In: *Physics for Diagnostic Radiology: Third Edition*. ; 2011. doi:10.1017/s0266462300001501

6. Shukla A, Kumar U. Positron emission tomography: An overview. *J Med Phys*. 2006. doi:10.4103/0971-6203.25665

7. Kalender WA. X-ray computed tomography. *Phys Med Biol*. 2006. doi:10.1088/0031-9155/51/13/R03

8. Geise RA. Computed Tomography: Physical Principles, Clinical Applications, and Quality Control. *Radiology*. 1995. doi:10.1148/radiology.194.3.782

9. Jane Dobbs H, Parker RP, Hodson NJ, Hobday P, Husband JE. The use of CT in radiotherapy treatment planning. *Radiother Oncol*. 1983;1(2). doi:10.1016/S0167-8140(83)80016-4

10. Pouliot J, Bani-Hashemi A, Josephine Chen, et al. Low-dose megavoltage cone-beam CT for radiation therapy. *Int J Radiat Oncol Biol Phys*. 2005;61(2):552-560. doi:10.1016/j.ijrobp.2004.10.011

11. Wang J, Li T, Liang Z, Xing L. Dose reduction for kilovotage cone-beam computed tomography in radiation therapy. *Phys Med Biol*. 2008;53(11):2897-2909. doi:10.1088/0031-9155/53/11/009

12. Cormack AM. Representation of a function by its line integrals, with some radiological applications. II. *J Appl Phys*. 1964. doi:10.1063/1.1713127

13. Oldendorf WH. Isolated Flying Spot Detection of Radiodensity Discontinuities—Displaying the Internal Structural Pattern of a Complex Object. *Ire Trans Biomed Electron*. 1961. doi:10.1109/TBMEL.1961.4322854

14. Agrawal MD, Pinho DF, Kulkarni NM, Hahn PF, Guimaraes AR, Sahani D V. Oncologic Applications of Dual-Energy CT in the Abdomen. *RadioGraphics*. 2014;34(3):589-612. doi:10.1148/rg.343135041

15. Lee HA, Lee YH, Yoon KH, Bang DH, Park DE. Comparison of virtual unenhanced images derived from dual-energy CT with true unenhanced images in evaluation of gallstone disease. *Am J Roentgenol*. 2016. doi:10.2214/AJR.15.14570

16. Wang Q, Shi G, Qi X, Fan X, Wang L. Quantitative analysis of the dual-energy CT virtual spectral curve for focal liver lesions characterization. *Eur J Radiol*. 2014. doi:10.1016/j.ejrad.2014.07.009

17. Boll DT, Patil NA, Paulson EK, et al. Renal Stone Assessment with Dual-Energy Multidetector CT and Advanced Postprocessing Techniques: Improved Characterization of Renal Stone Composition—Pilot Study. *Radiology*. 2009. doi:10.1148/radiol.2503080545

18. De Cecco CN, Darnell A, Rengo M, et al. Dual-Energy CT: Oncologic Applications. *Am J Roentgenol*. 2012;199(5_supplement):S98-S105. doi:10.2214/AJR.12.9207

19. Alvarez RE, MacOvski A. Energy-selective reconstructions in X-ray computerised tomography. *Phys Med Biol*. 1976. doi:10.1088/0031-9155/21/5/002

20. Macovski A, Alvarez RE, Chan JLH, Stonestrom JP, Zatz LM. Energy dependent reconstruction in X-ray computerized tomography. *Comput Biol Med*. 1976. doi:10.1016/0010-4825(76)90069-X

21. Mendonça PRS, Lamb P, Sahani D V. A Flexible Method for Multi-Material Decomposition of Dual-Energy CT Images. *IEEE Trans Med Imaging*. 2014;33(1):99-116. doi:10.1109/TMI.2013.2281719

22. Otto K. Volumetric modulated arc therapy: IMRT in a single gantry arc. *Med Phys*. 2007;35(1):310-317. doi:10.1118/1.2818738

23. Yu CX. Intensity-modulated arc therapy with dynamic multileaf collimation: an alternative to tomotherapy. *Phys Med Biol*. 1995;40(9):1435. doi:10.1088/0031-9155/40/9/004

24. Brahme A, Roos JE, Lax I. Solution of an integral equation encountered in rotation therapy. *Phys Med Biol*. 1982. doi:10.1088/0031-9155/27/10/002

25. Wright SJ. *Primal-Dual Interior-Point Methods*. SIAM; 1997.

26. Parikh N, Boyd S. Proximal Algorithms. *Found Trends® Optim*. 2014;1(3):127-239. doi:10.1561/2400000003

27. Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm. *SIAM J Imaging Sci*. 2009;2(1):183. doi:10.1137/080716542

28. Boyd S. Alternating Direction Method of Multipliers. *Proc 51st IEEE Conf Decis Control*. 2011;3(1):1-44. doi:10.1561/2200000016

29. Chambolle A, Pock T. A first-order primal-dual algorithm for convex problems with applications to imaging. *J Math Imaging Vis*. 2011;40(1):120-145. doi:10.1007/s10851-010-0251-1

30. Lyu Q, Yang C, Gao H, et al. Technical Note: Iterative megavoltage CT (MVCT) reconstruction using block-matching 3D-transform (BM3D) regularization. *Med Phys*. 2018;45(6):2603-2610. doi:10.1002/mp.12916

31. Lyu Q, Ruan D, Hoffman J, Neph R, McNitt-Gray M, Sheng K. Iterative reconstruction for low dose CT using Plug-and-Play alternating direction method of multipliers (ADMM) framework. In: *Proc.SPIE*. Vol 10949. ; 2019. https://doi.org/10.1117/12.2512484.

32. Lyu Q, Niu T, Ruan D, Sheng K. Iterative reconstruction of cone-beam breast {CT} using plug-and-play projected gradient descent. In: *Medical Imaging 2020: Physics of Medical Imaging*. Vol 11312. International Society for Optics and Photonics; 2020:113124J.

33. Lyu Q, O'Connor D, Niu T, Sheng K. Image-domain multi-material decomposition for dual-energy CT with non-convex sparsity regularization. In: *Proc.SPIE*. Vol 10949. ; 2019. https://doi.org/10.1117/12.2508037.

34. Lyu Q, O'Connor D, Niu T, Sheng K. Image-domain multimaterial decomposition for dual-energy computed tomography with nonconvex sparsity regularization. *J Med Imaging*. 2019;6(4):44004.

35. Nguyen D, Lyu Q, Ruan D, et al. A comprehensive formulation for volumetric modulated arc therapy planning. *Med Phys*. 2016;43(7):4263-4272. doi:10.1118/1.4953832

36. Lyu Q, Neph R, Yu VY, Ruan D, Sheng K. Single-arc VMAT optimization for dual-layer MLC. *Phys Med Biol*. 2019;64(9). doi:10.1088/1361-6560/ab0ddd

37. Lyu Q, Neph R, Yu VY, Ruan D, Boucher S, Sheng K. Many-isocenter optimization for robotic radiotherapy. *Phys Med Biol*. 2020;65(4). doi:10.1088/1361-6560/ab63b8

38. Lyu Q, O'Connor D, Ruan D, Yu V, Nguyen D, Sheng K. VMAT optimization with dynamic collimator rotation. *Med Phys*. 2018;45(6). doi:10.1002/mp.12915

39. Lyu Q, Yu VY, Ruan D, Neph R, O'Connor D, Sheng K. A novel optimization framework for VMAT with dynamic gantry couch rotation. *Phys Med Biol*. 2018;63(12). doi:10.1088/1361-6560/aac704

40. Lyu Q, Neph R, O'Connor D, Ruan D, Boucher S, Sheng K. ROAD: ROtational direct Aperture optimization with a Decoupled ring-collimator for FLASH radiotherapy. *Phys Med Biol*. 2021;66(3). doi:10.1088/1361-6560/abcbd0

41. Beavis AW. Is tomotherapy the future of IMRT? *Br J Radiol*. 2004;77(916):285-295. doi:10.1259/bjr/22666727

42. Yartsev S, Kron T, Van Dyk J. Tomotherapy as a tool in image-guided radiation therapy (IGRT): theoretical and technological aspects. *Biomed Imaging Interv J*. 2007;3(1):e16. doi:10.2349/biij.3.1.e16

43. Bian J, Siewerdsen JH, Han X, et al. Evaluation of sparse-view reconstruction from flat-panel-detector cone-beam CT. *Phys Med Biol*. 2010;55(22):6575-6599. doi:10.1088/0031-9155/55/22/001

44. LaRoque SJ, Sidky EY, Pan X. Accurate image reconstruction from few-view and limited-angle data in diffraction tomography. *J Opt Soc Am A Opt Image Sci Vis*. 2008;25(7):1772-1782. doi:10.1364/JOSAA.25.001772

45. Chen G-H, Tang J, Leng S. Prior image constrained compressed sensing (PICCS): A method to accurately reconstruct dynamic CT images from highly undersampled projection data sets. *Med Phys*. 2008;35(2):660-663. doi:10.1118/1.2836423

46. Tian Z, Jia X, Yuan K, Pan T, Jiang SB. Low-dose CT reconstruction via edge-preserving total variation regularization. *Phys Med Biol*. 2011;56(18):5949-5967. doi:10.1088/0031-9155/56/18/011

47. Gao H, Qi XS, Gao Y, Low D a. Megavoltage CT imaging quality improvement on TomoTherapy via tensor framelet. *Med Phys*. 2013;40(8):081919. doi:10.1118/1.4816303

48. Lebrun M. An Analysis and Implementation of the BM3D Image Denoising Method. *Image Process Line*. 2012;2:175-213. doi:10.5201/ipol.2012.l-bm3d

49. Dabov K, Foi A, Katkovnik V. Image denoising by sparse 3D transformation-domain collaborative filtering. *IEEE Trans Image Process*. 2007;16(8):1-16. doi:10.1109/TIP.2007.901238

50. Huang J, Ma J, Liu N, Feng Q, Chen W. Projection data restoration guided non-local means for low-dose computed tomography reconstruction. In: *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. ; 2011:1167-1170. doi:10.1109/ISBI.2011.5872609

51. Sheng K, Gou S, Wu J, Qi SX. Denoised and texture enhanced MVCT to improve soft tissue conspicuity. *Med Phys*. 2014;41(10):101916. doi:10.1118/1.4894714

52. Trinh H, Luong M, Rocchisani J, Dibos F. *An Optimal Weight Method for CT Image Denoising*. Vol 10.; 2012.

53. Kang D, Slomka P, Nakazato R, et al. *Image Denoising of Low-Radiation Dose Coronary CT Angiography by an Adaptive Block-Matching 3D Algorithm*. Vol 8669.; 2013. doi:10.1117/12.2006907

54. Danielyan A, Katkovnik V, Egiazarian K. BM3D frames and variational image deblurring. *IEEE Trans Image Process*. 2012;21(4):1715-1728. doi:10.1109/TIP.2011.2176954

55. Katkovnik V, Danielyan A, Egiazarian K. Decoupled inverse and denoising for image deblurring: Variational BM3D-frame technique. *Proc - Int Conf Image Process ICIP*. 2011:3453-3456. doi:10.1109/ICIP.2011.6116455

56. Danielyan A, Katkovnik V, Egiazarian K. Image Deblurring By Augmented Langrangian With Bm3D Frame Prior. *Third Work Inf Theor Methods Sci Eng*. 2010.

http://dspace.cc.tut.fi/dpub/handle/123456789/8037?show=full.

57. Yang L, Getreuer P, Linghong Z. Sparse-view cone-beam CT reconstruction via previous normal dose scan induced BM3D-frame regularization method. In: *The 12th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*. ; 2013:537-540.

58. Gayou O, Miften M. Commissioning and clinical implementation of a mega-voltage cone beam CT system for treatment localization. *Med Phys*. 2007;34(8):3183-3192. doi:10.1118/1.2752374

59. Moore KL, Palaniswaamy G, White B, Goddu SM, Low D a. Fast, low-dose patient localization on TomoTherapy via topogram registration. *Med Phys*. 2010;37(8):4068-4077. doi:10.1118/1.3453577

60. Richard S, Husarik DB, Yadava G, Murphy SN, Samei E. Towards task-based assessment of CT performance: system and object MTF across different reconstruction algorithms. *Med Phys*. 2012;39(7Part1):4115-4122. doi:10.1118/1.4725171

61. Baskaran MM. Optimizing Sparse Matrix-Vector Multiplication on GPUs using Compile-time and Run-time Strategies. *IBM Res Rep*. 2008;24704. doi:10.1145/1366230.1366244

62. Greffier J, Fernandez A, Macri F, Freitag C, Metge L, Beregi JP. Which dose for what image? Iterative reconstruction for CT scan. *Diagn Interv Imaging*. 2013. doi:10.1016/j.diii.2013.03.008

63. Qihui L, Chunlin Y, Hao G, et al. Technical Note: Iterative megavoltage CT (MVCT) reconstruction using block-matching 3D-transform (BM3D) regularization. *Med Phys*.

2018;45(6):2603-2610. doi:10.1002/mp.12916

64. Venkatakrishnan S V, Bouman CA, Wohlberg B. Plug-and-Play priors for model based reconstruction. In: *2013 IEEE Global Conference on Signal and Information Processing*. ; 2013:945-948. doi:10.1109/GlobalSIP.2013.6737048

65. Chan SH, Wang X, Elgendy OA. Plug-and-Play ADMM for Image Restoration: Fixed Point Convergence and Applications. 2016;3(1):84-98. doi:10.1109/TCI.2016.2629286

66. Wang X, Chan SH. Parameter-free Plug-and-Play ADMM for image restoration. *ICASSP, IEEE Int Conf Acoust Speech Signal Process - Proc*. 2017:1323-1327. doi:10.1109/ICASSP.2017.7952371

67. Sreehari S, Venkatakrishnan S V, Wohlberg B, et al. Plug-and-Play Priors for Bright Field Electron Tomography and Sparse Interpolation. *IEEE Trans Comput Imaging*. 2016;2(4):408-423. doi:10.1109/TCI.2016.2599778

68. Kershaw DS. The incomplete Cholesky—conjugate gradient method for the iterative solution of systems of linear equations. *J Comput Phys*. 1978;26(1):43-65. doi:https://doi.org/10.1016/0021-9991(78)90098-0

69. Kelley CT. Iterative Optimization. *Society*. 1999. doi:10.1007/978-3-540-78879-9

70. Hoffman J, Young S, Noo F, McNitt-Gray M. Technical Note: FreeCT-wFBP: A robust, efficient, open-source implementation of weighted filtered backprojection for helical, fan-beam CT. *Med Phys*. 2016;43(3):1411-1420. doi:10.1118/1.4941953

71. Xie X, McGaffin MG, Long Y, Fessler JA, Wen M, Lin J. Accelerating separable footprint

(SF) forward and back projection on GPU. In: Vol 10132. ; 2017:101322S-10132-10138. https://doi.org/10.1117/12.2252010.

72. Lo P, Young S, Kim HJ, Brown MS, McNitt-Gray MF. Variability in CT lung-nodule quantification: Effects of dose reduction and reconstruction methods on density and texture based features: Effects. *Med Phys*. 2016;43(8):4854-4865. doi:10.1118/1.4954845

73. Žabić S, Wang Q, Morton T, Brown KM. A low dose simulation tool for CT systems with energy integrating detectors. In: *Medical Physics*. ; 2013. doi:10.1118/1.4789628

74. Zhang K, Zuo W, Gu S, Zhang L. Learning deep CNN denoiser prior for image restoration. In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. ; 2017. doi:10.1109/CVPR.2017.300

75. Owens JD, Houston M, Luebke D, Green S, Stone JE, Phillips JC. GPU computing. *Proc IEEE*. 2008. doi:10.1109/JPROC.2008.917757

76. Ono S. Primal-Dual Plug-and-Play Image Restoration. *IEEE Signal Process Lett*. 2017;24(8):1108-1112. doi:10.1109/LSP.2017.2710233

77. Kamilov US, Mansour H, Wohlberg B. A Plug-and-Play Priors Approach for Solving Nonlinear Imaging Inverse Problems. *IEEE Signal Process Lett*. 2017;24(12):1872-1876. doi:10.1109/LSP.2017.2763583

78. Chambolle-pock T. Preconditioned Chambolle-Pock : The block separable. :1-2.

79. Jin KH, McCann MT, Froustey E, Unser M. Deep Convolutional Neural Network for Inverse Problems in Imaging. *IEEE Trans Image Process*. 2017.

doi:10.1109/TIP.2017.2713099

80. Lyu Q, Yang C, Gao H, et al. Technical Note: Iterative megavoltage CT (MVCT) reconstruction using block-matching 3D-transform (BM3D) regularization. *Med Phys*. 2018. doi:10.1002/mp.12916

81. Gupta H, Jin KH, Nguyen HQ, McCann MT, Unser M. CNN-Based Projected Gradient Descent for Consistent CT Image Reconstruction. *IEEE Trans Med Imaging*. 2018. doi:10.1109/TMI.2018.2832656

82. Gao H. Fast parallel algorithms for the x-ray transform and its adjoint. *Med Phys*. 2012;39(11):7110-7120. doi:10.1118/1.4761867

83. McCollough CH, Leng S, Yu L, Fletcher JG. Dual- and Multi-Energy CT: Principles, Technical Approaches, and Clinical Applications. *Radiology*. 2015. doi:10.1148/radiol.2015142631

84. Long Y, Fessler JA. Multi-material decomposition using statistical image reconstruction for spectral CT. *IEEE Trans Med Imaging*. 2014;33(8):1614-1626. doi:10.1109/TMI.2014.2320284

85. Moellenhoff T, Strekalovskiy E, Möller M, Cremers D. Low Rank Priors for Color Image Regularization. *Submitted*. 2014:1-14. doi:10.1007/978-3-319-14612-6_10

86. Malitsky Y, Pock T. A first-order primal-dual algorithm with linesearch. *ArXiv e-prints*. August 2016.

87. Wang W, Carreira-Perpiñán M ~Á. ~Á. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *ArXiv e-prints*. September

2013.

88. Boyd S, Vandenberghe L. *Convex Optimization*. Vol 25.; 2010. doi:10.1080/10556781003625177

89. Pock T, Cremers D, Bischof H, Chambolle A. An algorithm for minimizing the Mumford-Shah functional. In: *Proceedings of the IEEE International Conference on Computer Vision*. ; 2009. doi:10.1109/ICCV.2009.5459348

90. Earl MA, Shepard DM, Naqvi S, Li XA, Yu CX. Inverse planning for intensity-modulated are therapy using direct aperture optimization. *Phys Med Biol*. 2003. doi:10.1088/0031-9155/48/8/309

91. Wong E, Chen JZ, Greenland J. Intensity-modulated ARC therapy simplified. *Int J Radiat Oncol Biol Phys*. 2002. doi:10.1016/S0360-3016(02)02735-9

92. Crooks SM, Wu X, Takita C, Watzich M, Xing L. Aperture modulated arc therapy. *Phys Med Biol*. 2003. doi:10.1088/0031-9155/48/10/307

93. Bratengeier K, Gainey M, Sauer OA, Richter A, Flentje M. Fast intensity-modulated arc therapy based on 2-step beam segmentation. *Med Phys*. 2011. doi:10.1118/1.3523602

94. Locke CB, Bush KK. Trajectory optimization in radiotherapy using sectioning (TORUS): *Med Phys*. 2017;44(7):3375-3392. doi:10.1002/mp.12270

95. Liu Y, Shi C, Lin B, Ha CS, Papanikolaou N. Delivery of four-dimensional radiotherapy with TrackBeam for moving target using a dual-layer MLC: Dynamic phantoms study. *J Appl Clin Med Phys*. 2009. doi:10.1120/jacmp.v10i2.2926

96. Giantsoudi D, Stathakis S, Liu Y, Shi C, Papanikolaou N. Monte Carlo modeling and

commissioning of a dual-layer micro multileaf collimator. *Technol Cancer Res Treat*. 2009. doi:10.1177/153303460900800203

97. Liu Y, Shi C, Tynan P, Papanikolau N. Dosimetric characteristics of dual-layer multileaf collimation for small-field and intensity-modulated radiation therapy applications. *J Appl Clin Med Phys*. 2008;9(2):15-29. doi:10.1120/jacmp.v9i2.2709

98. Michiels S, Poels K, Crijns W, et al. Volumetric modulated arc therapy of head-and-neck cancer on a fast-rotating O-ring linac: Plan quality and delivery time comparison with a C-arm linac. *Radiother Oncol*. 2018;128(3):479-484. doi:10.1016/J.RADONC.2018.04.021

99. Shaverdian N, Yang Y, Hu P, et al. Feasibility evaluation of diffusion-weighted imaging using an integrated MRI-radiotherapy system for response assessment to neoadjuvant therapy in rectal cancer. *Br J Radiol*. 2017. doi:10.1259/bjr.20160739

100. Neylon J, Sheng K, Yu V, et al. A nonvoxel-based dose convolution/superposition algorithm optimized for scalable GPU architectures. *Med Phys*. 2014;41(10):101711. doi:10.1118/1.4895822

101. Grégoire V, Mackie TR. State of the art on dose prescription, reporting and recording in intensity-modulated radiation therapy (ICRU report No. 83). *Cancer/Radiotherapie*. 2011;15(6-7):555. doi:10.1016/j.canrad.2011.04.003

102. Riley C, Cox C, Graham S, et al. Varian Halcyon dosimetric comparison for multiarc VMAT prostate and head-and-neck cancers. *Med Dosim*. September 2018. doi:10.1016/J.MEDDOS.2018.06.004

103. Court L. Experience and Suggestions on “ Commissioning and QA Procedures ” for

Halcyon.

104. Rao M, Yang W, Chen F, et al. Comparison of Elekta VMAT with helical tomotherapy and fixed field IMRT: plan quality, delivery efficiency and accuracy. *Med Phys*. 2010;37(3):1350. doi:10.1118/1.3326965

105. Teoh M, Clark CH, Wood K, Whitaker S, Nisbet A. Volumetric modulated arc therapy: A review of current literature and clinical use in practice. *Br J Radiol*. 2011. doi:10.1259/bjr/22373346

106. Vanetti E, Clivio A, Nicolini G, et al. Volumetric modulated arc radiotherapy for carcinomas of the oro-pharynx, hypo-pharynx and larynx: A treatment planning comparison with fixed field IMRT. *Radiother Oncol*. 2009. doi:10.1016/j.radonc.2008.12.008

107. Dong P, Lee P, Ruan D, et al. 4π noncoplanar stereotactic body radiation therapy for centrally located or larger lung tumors. *Int J Radiat Oncol Biol Phys*. 2013;86(3):407. doi:10.1016/j.ijrobp.2013.02.002

108. Dong P, Lee P, Ruan D, et al. 4π non-coplanar liver SBRT: a novel delivery technique. *Int J Radiat Oncol Biol Phys*. 2013;85:1360. doi:10.1016/j.ijrobp.2012.09.028

109. Smyth G, Evans PM, Bamber JC, Bedford JL. Recent developments in non-coplanar radiotherapy. *Br J Radiol*. 2019;92(1097):20180908. doi:10.1259/bjr.20180908

110. Meedt G, Alber M, Nüsslin F. Non-coplanar beam direction optimization for intensity-modulated radiotherapy. *Phys Med Biol*. 2003. doi:10.1088/0031-9155/48/18/304

111. Laing R, Bentley R, Nahum A, Brada M. Stereotactic radiotherapy for irregular targets:

Comparison between non-coplanar arcs and static non-coplanar conformal beams. *Eur J Cancer*. 1993. doi:10.1016/0959-8049(93)91857-h

112. Chapet O, Khodri M, Jalade P, et al. Potential benefits of using non coplanar field and intensity modulated radiation therapy to preserve the heart in irradiation of lung tumors in the middle and lower lobes. *Radiother Oncol*. 2006. doi:10.1016/j.radonc.2006.07.009

113. Smyth G, Bamber JC, Evans PM, Bedford JL. Trajectory optimization for dynamic couch rotation during volumetric modulated arc radiotherapy. *Phys Med Biol*. 2013;58(22):8163. doi:10.1088/0031-9155/58/22/8163

114. Murzin VL, Woods K, Moiseenko V, et al. 4π plan optimization for cortical-sparing brain radiotherapy. *Radiother Oncol*. 2018. doi:10.1016/j.radonc.2018.02.011

115. Dong P, Yu V, Nguyen D, et al. Feasibility of using intermediate x-ray energies for highly conformal extracranial radiotherapy. *Med Phys*. 2014. doi:10.1118/1.4868464

116. Pugachev A, Li JG, Boyer AL, et al. Role of beam orientation optimization in intensity-modulated radiation therapy. *Int J Radiat Oncol Biol Phys*. 2001;50(2):551-560. doi:10.1016/S0360-3016(01)01502-4

117. Voet PWJ, Breedveld S, Dirkx MLP, Levendag PC, Heijmen BJM. Integrated multicriterial optimization of beam angles and intensity profiles for coplanar and noncoplanar head and neck IMRT and implications for VMAT. *Med Phys*. 2012. doi:10.1118/1.4736803

118. Krayenbuehl J, Davis JB, Ciernik IF. Dynamic intensity-modulated non-coplanar arc radiotherapy (INCA) for head and neck cancer. *Radiother Oncol*. 2006;81(2):151.

doi:10.1016/j.radonc.2006.09.004

119. Yang J, Fowler JF, Lamond JP, Lanciano R, Feng J, Brady LW. Red Shell: Defining a High-Risk Zone of Normal Tissue Damage in Stereotactic Body Radiation Therapy. *Int J Radiat Oncol Biol Phys*. 2010. doi:10.1016/j.ijrobp.2009.12.069

120. Yu VY, Landers A, Woods K, et al. A prospective 4π radiation therapy clinical study in recurrent high-grade glioma patients. *Int J Radiat Oncol Biol Phys*. 2018;101(1):144. doi:10.1016/j.ijrobp.2018.01.048

121. Smyth G, Evans PM, Bamber JC, et al. Dosimetric accuracy of dynamic couch rotation during volumetric modulated arc therapy (DCR-VMAT) for primary brain tumours. *Phys Med Biol*. 2019. doi:10.1088/1361-6560/ab0a8e

122. Wilson B, Otto K, Gete E. A simple and robust trajectory-based stereotactic radiosurgery treatment: *Med Phys*. 2017. doi:10.1002/mp.12036

123. Pollom E, Wang L, Gibbs IC, Soltys SG. CyberKnife Robotic Stereotactic Radiosurgery BT - Stereotactic Radiosurgery and Stereotactic Body Radiation Therapy: A Comprehensive Guide. In: Trifiletti DM, Chao ST, Sahgal A, Sheehan JP, eds. Cham: Springer International Publishing; 2019:67-76. doi:10.1007/978-3-030-16924-4_6

124. Dong P, Nguyen D, Ruan D, et al. Feasibility of prostate robotic radiation therapy on conventional C-arm linacs. *Pract Radiat Oncol*. 2014. doi:10.1016/j.prro.2013.10.009

125. Kim MM, Kennedy C, Scheuermann R, Freedman G, Li T. Whole Breast and Lymph Node Irradiation using Halcyon™ 2.0 Utilizing Automatic Multi-isocenter Treatment Delivery and Daily Kilovoltage Cone-beam Computed Tomography. *Cureus*. 2019. doi:10.7759/cureus.4744

126. Amoush A, Murray E, Yu JS, Xia P. Single-isocenter hybrid IMRT plans versus two-isocenter conventional plans and impact of intrafraction motion for the treatment of breast cancer with supraclavicular lymph nodes involvement. *J Appl Clin Med Phys*. 2015. doi:10.1120/jacmp.v16i4.5188

127. Wang Z, Jiang W, Feng Y, et al. A simple approach of three-isocenter IMRT planning for craniospinal irradiation. *Radiat Oncol*. 2013. doi:10.1186/1748-717X-8-217

128. Hong L, Alektiar K, Chui C, et al. IMRT of large fields: Whole-abdomen irradiation. *Int J Radiat Oncol Biol Phys*. 2002. doi:10.1016/S0360-3016(02)02921-8

129. Zeng GG, Heaton RK, Catton CN, et al. A two isocenter IMRT technique with a controlled junction dose for long volume targets. *Phys Med Biol*. 2007. doi:10.1088/0031-9155/52/15/012

130. Zhu L, Lee L, Ma Y, Ye Y, Mazzeo R, Xing L. Using total-variation regularization for intensity modulated radiation therapy inverse planning with field-specific numbers of segments. *Phys Med Biol*. 2008;53(23):6653. doi:10.1088/0031-9155/53/23/002

131. Blumensath T, Davies ME. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE J Sel Top Signal Process*. 2010. doi:10.1109/JSTSP.2010.2042411

132. Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J Imaging Sci*. 2009;2(1):183. doi:10.1137/080716542

133. Yu VY, Tran A, Nguyen D, et al. The development and verification of a highly accurate collision prediction model for automated noncoplanar plan delivery. *Med Phys*. 2015;42(11):6457-6467. doi:10.1118/1.4932631

134. Neph R, Ouyang C, Neylon J, Yang Y, Sheng K. Parallel beamlet dose calculation via beamlet contexts in a distributed multi-GPU framework. *Med Phys*. 2019;0(0). doi:doi:10.1002/mp.13651

135. Xia P, Verhey LJ. Multileaf collimator leaf sequencing algorithm for intensity modulated beams with multiple static segments. *Med Phys*. 1998. doi:10.1118/1.598315

136. Reinelt G. TSPLIB—A Traveling Salesman Problem Library. *ORSA J Comput*. 1991. doi:10.1287/ijoc.3.4.376

137. Nguyen D, Rwigema J-CM, Yu VY, et al. Feasibility of extreme dose escalation for glioblastoma multiforme using 4pi radiotherapy. *Radiat Oncol*. 2014;9(1):239. doi:10.1186/s13014-014-0239-x

138. Breedveld S, Storchi PRM, Voet PWJ, Heijmen BJM. ICycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans. *Med Phys*. 2012. doi:10.1118/1.3676689

139. Smyth G, Evans PM, Bamber JC, et al. Non-coplanar trajectories to improve organ at risk sparing in volumetric modulated arc therapy for primary brain tumors. *Radiother Oncol*. 2016;121(1):124-131. doi:10.1016/j.radonc.2016.07.014

140. Lin YW, Lin KH, Ho HW, et al. Treatment plan comparison between stereotactic body radiation therapy techniques for prostate cancer: Non-isocentric CyberKnife versus isocentric RapidArc. *Phys Medica*. 2014. doi:10.1016/j.ejmp.2014.03.008

141. O'Connor D, Voronenko Y, Nguyen D, Yin W, Sheng K. Fast non-coplanar beam orientation optimization based on group sparsity. October 2017.

https://arxiv.org/abs/1710.05308. Accessed July 25, 2019.

142. Li Y, Yao J, Yao D. Automatic beam angle selection in IMRT planning using genetic algorithm. *Phys Med Biol*. 2004;49(10):1915-1932. doi:10.1088/0031-9155/49/10/007

143. Li Y, Yao D, Yao J, Chen W. A particle swarm optimization algorithm for beam angle selection in intensity-modulated radiotherapy planning. *Phys Med Biol*. 2005;50(15):3491-3514. doi:10.1088/0031-9155/50/15/002

144. Pugachev A, Xing L. Incorporating prior knowledge into beam orientaton optimization in IMRT. *Int J Radiat Oncol Biol Phys*. 2002. doi:10.1016/S0360-3016(02)03917-2

145. Romeijn HE, Ahuja RK, Dempsey JF, Kumar A. A Column Generation Approach to Radiation Therapy Treatment Planning Using Aperture Modulation. *SIAM J Optim*. 2005. doi:10.1137/040606612

146. Tran A, Zhang J, Woods K, et al. Treatment planning comparison of IMPT, VMAT and 4π radiotherapy for prostate cases. *Radiat Oncol*. 2017;12(1):10. doi:10.1186/s13014-016-0761-0

147. Woods K, Lee P, Kaprealian T, Yang I, Sheng K. Cochlea-sparing acoustic neuroma treatment with 4π radiation therapy. *Adv Radiat Oncol*. 2018. doi:10.1016/j.adro.2018.01.004

148. Woods K, Nguyen D, Tran A, et al. Viability of Noncoplanar VMAT for liver SBRT compared with coplanar VMAT and beam orientation optimized 4π IMRT. *Adv Radiat Oncol*. 2016;1(1):67-75. doi:10.1016/j.adro.2015.12.004

149. Bortfeld T, Webb S. Single-Arc IMRT? *Phys Med Biol*. 2009;54(1):N9-N20. doi:10.1088/0031-9155/54/1/N02

150. Tas B, Bilge H, Ozturk ST. An investigation of the dose distribution effect related with collimator angle in volumetric arc therapy of prostate cancer. *J Med Phys*. 2016;41(2):100-105. doi:10.4103/0971-6203.181635

151. Ahn BS, Park S-Y, Park JM, Choi CH, Chun M, Kim J. Dosimetric effects of sectional adjustments of collimator angles on volumetric modulated arc therapy for irregularly-shaped targets. *PLoS One*. 2017;12(4):1-14. doi:10.1371/journal.pone.0174924

152. Zhang P, Happersett L, Yang Y, Yamada Y, Mageras G, Hunt M. Optimization of Collimator Trajectory in Volumetric Modulated Arc Therapy: Development and Evaluation for Paraspinal SBRT. *Int J Radiat Oncol*. 2010;77(2):591-599. doi:10.1016/j.ijrobp.2009.08.056

153. Kim H, Li R, Lee R, Xing L. Beam's-eye-view dosimetrics (BEVD) guided rotational station parameter optimized radiation therapy (SPORT) planning based on reweighted total-variation minimization. *Phys Med Biol*. 2015;60(5):N71-N82. doi:10.1088/0031-9155/60/5/N71

154. Bollobás B, Riordan O. Dijkstra ' s Algorithm. *Network*. 1993;69(1959):036114. http://www.ncbi.nlm.nih.gov/pubmed/21282851.

155. Ee236C. Fast proximal gradient methods. 2013:1-32.

156. Noe M. The Annals of Mathematical Statistics. *Ann Math Stat*. 1930;1(1):1-2. doi:10.1214/aoms/1177733256

157. Sharfo AWM, Dirkx MLP, Breedveld S, Romero AM, Heijmen BJM. VMAT plus a few computer-optimized non-coplanar IMRT beams (VMAT+) tested for liver SBRT. *Radiother Oncol*. 2017;123(1):49. doi:10.1016/j.radonc.2017.02.018

158. Shaitelman SF, Kim LH, Yan D, Martinez AA, Vicini FA, Grills IS. Continuous arc rotation of the couch therapy for the delivery of accelerated partial breast irradiation: a treatment planning analysis. *Int J Radiat Oncol Biol Phys*. 2011;80(3):771. doi:10.1016/j.ijrobp.2010.03.004

159. Liang J, Atwood T, Von Eyben R, et al. Trajectory modulated arc therapy: a fully dynamic delivery with synchronized couch and gantry motion significantly improves dosimetric indices correlated with poor cosmesis in accelerated partial breast irradiation. *Int J Radiat Oncol Biol Phys*. 2015;92(5):1148. doi:10.1016/j.ijrobp.2015.04.034

160. Yang Y, Zhang P, Happersett L, et al. Choreographing couch and collimator in volumetric modulated arc therapy. *Int J Radiat Oncol Biol Phys*. 2011;80(4):1238. doi:10.1016/j.ijrobp.2010.10.016

161. Lee MacDonald R, Thomas CG. Dynamic trajectory-based couch motion for improvement of radiation therapy trajectories in cranial SRT. *Med Phys*. 2015;42(5):2317. doi:10.1118/1.4917165

162. Wild E, Bangert M, Nill S, Oelfke U. Noncoplanar VMAT for nasopharyngeal tumors: plan quality versus treatment time. *Med Phys*. 2015;42(5):2157. doi:10.1118/1.4914863

163. Papp D, Bortfeld T, Unkelbach J. A modular approach to intensity-modulated arc

therapy optimization with noncoplanar trajectories. *Phys Med Biol*. 2015;60(13):5179. doi:10.1088/0031-9155/60/13/5179

164. Dijkstra EW. A note on two problems in connexion with graphs. *Numer Math*. 1959;1(1):269-271. doi:10.1007/BF01386390

165. Chang JY, Balter PA, Dong L, et al. Stereotactic body radiation therapy in centrally and superiorly located stage I or isolated recurrent non-small-cell lung cancer. *Int J Radiat Oncol Biol Phys*. 2008;72(4):967. doi:10.1016/j.ijrobp.2008.08.001

166. King CR, Brooks JD, Gill H, Presti JC. Long-term outcomes from a prospective trial of stereotactic body radiotherapy for low-risk prostate cancer. *Int J Radiat Oncol Biol Phys*. 2012;82(2):877-882. doi:10.1016/j.ijrobp.2010.11.054

167. Ohira S, Ueda Y, Akino Y, et al. HyperArc VMAT planning for single and multiple brain metastases stereotactic radiosurgery: a new treatment planning approach. *Radiat Oncol*. 2018;13(1):13. doi:10.1186/s13014-017-0948-z

168. Bortfeld T. IMRT: A review and preview. *Phys Med Biol*. 2006. doi:10.1088/0031-9155/51/13/R21

169. Palma DA, Verbakel WFAR, Otto K, Senan S. New developments in arc radiation therapy: A review. *Cancer Treat Rev*. 2010. doi:10.1016/j.ctrv.2010.01.004

170. Favaudon V, Caplier L, Monceau V, et al. Ultrahigh dose-rate FLASH irradiation increases thedifferential response between normal and tumor tissue in mice. *Sci Transl Med*. 2014;6(245):245ra93.

171. Favaudon V, Caplier L, Monceau V, et al. Ultrahigh dose-rate FLASH irradiation

increases the differential response between normal and tumor tissue in mice. *Sci Transl Med*. 2014. doi:10.1126/scitranslmed.3008973

172. Vozenin M-C, De Fornel P, Petersson K, et al. The Advantage of FLASH Radiotherapy Confirmed in Mini-pig andCat-cancer Patients. *Clin Cancer Res*. 2019;25(1):35-42.

173. Vozenin M-C, Hendry JH, Limoli CL. Biological Benefits of Ultra-high Dose Rate FLASH Radiotherapy:Sleeping Beauty Awoken. *Clin Oncol*. 2019;31(7):407-415.

174. Bourhis J, Montay-Gruel P, Gonçalves Jorge P, et al. Clinical translation of FLASH radiotherapy: Why and how? *Radiother Oncol*. 2019;139:11-17.

175. Beyreuther E, Brand M, Hans S, et al. Feasibility of proton FLASH effect tested by zebrafish embryoirradiation. *Radiother Oncol*. 2019;139:46-50.

176. Montay-Gruel P, Petersson K, Jaccard M, et al. Irradiation in a flash: Unique sparing of memory in mice after whole brain irradiation with dose rates above 100 Gy/s. *Radiother Oncol*. 2017. doi:10.1016/j.radonc.2017.05.003

177. Simmons DA, Lartey FM, Schüler Emiland Rafat M, et al. Reduced cognitive deficits after FLASH irradiation of wholemouse brain are associated with less hippocampal dendritic spineloss and neuroinflammation. *Radiother Oncol*. 2019;139:4-10.

178. Lempart M, Blad B, Adrian G, et al. Modifying a clinical linear accelerator for delivery ofultra-high dose rate irradiation. *Radiother Oncol*. 2019;139:40-45.

179. Montay-Gruel P, Bouchet A, Jaccard M, et al. X-rays can trigger the FLASH effect: Ultra-high dose-ratesynchrotron light source prevents normal brain injury after wholebrain irradiation in mice. *Radiother Oncol*. 2018;129(3):582-588.

180. Bourhis J, Sozzi WJ, Jorge PGG calves, et al. Treatment of a first patient with FLASH-radiotherapy. *Radiother Oncol*. 2019;139:18-22. doi:10.1016/j.radonc.2019.06.019

181. Hamm RW, Hamm ME. *Industrial Accelerators and Their Applications.*; 2012. doi:10.1142/7745

182. Schwartz RB. NCRP Report No. 51, Radiation protection design guidelines for 0.1–100 MeV particle accelerator facilities. *Int J Appl Radiat Isot*. 1978. doi:10.1016/0020-708x(78)90087-x

183. Maxim PG, Tantawi SG, Loo Jr BW. PHASER: A platform for clinical translation of FLASH cancerradiotherapy. *Radiother Oncol*. 2019;139:28-33.

184. Lyu Q, Neph R, Yu VY, Ruan D, Sheng K. Single-arc VMAT optimization for dual-layer MLC. *Phys Med Biol*. 2019;64(9):095028. doi:10.1088/1361-6560/ab0ddd

185. Pratx G, Kapp DS. A computational model of radiolytic oxygen depletion during {FLASH} irradiation and its effect on the oxygen enhancement ratio. *Phys Med Biol*. 2019;64(18):185005. doi:10.1088/1361-6560/ab3769

186. Spitz DR, Buettner GR, Petronek Michael Sand St-Aubin JJ, Flynn RT, Waldron Timothy Jand Limoli CL. An integrated physico-chemical approach for explaining thedifferential impact of FLASH versus conventional dose rateirradiation on cancer and normal tissue responses. *Radiother Oncol*. 2019;139:23-27.

187. Michaels HB. Oxygen depletion in irradiated aqueous solutions containing electron affinic hypoxic cell radiosensitizers. *Int J Radiat Oncol Biol Phys*. 1986. doi:10.1016/0360-3016(86)90224-5

188. Montay-Gruel P, Acharya MM, Petersson K, et al. Long-term neurocognitive benefits of FLASH radiotherapy drivenby reduced reactive oxygen species. *Proc Natl Acad Sci U S A*. 2019;116(22):10943-10951.

189. Hendry JH, Moore J V., Hodgson BW, Keene JP. The Constant Low Oxygen Concentration in All the Target Cells for Mouse Tail Radionecrosis. *Radiat Res*. 1982;92(1):172. doi:10.2307/3575852

190. Pratx G, Kapp DS. Ultra-High-Dose-Rate FLASH Irradiation May Spare Hypoxic StemCell Niches in Normal Tissues. *Int J Radiat Oncol Biol Phys*. 2019;105(1):190-192.

191. Weiss H, Epp ER, Heslin JM, Ling CC, Santomasso A. Oxygen depletion in cells irradiated at ultra-high dose-rates and at conventional dose-rates. *Int J Radiat Biol*. 1974. doi:10.1080/09553007414550901

192. Ferrell RT, Himmelblau DM. Diffusion Coefficients of Nitrogen and Oxygen in Water. *J Chem Eng Data*. 1967. doi:10.1021/je60032a036

193. Hendry J. Taking Care with FLASH Radiation Therapy. *Int J Radiat Oncol Biol Phys*. 2020. doi:10.1016/j.ijrobp.2020.01.029

194. Fan Q, Nanduri A, Mazin S, Zhu L. Emission guided radiation therapy for lung and prostate cancers: A feasibility study on a digital patient. *Med Phys*. 2012. doi:10.1118/1.4761951