# Multicriteria VMAT optimization

David Craft, Dualta McQuaid, Jeremiah Wala, Wei Chen, Thomas Bortfeld

May 18, 2011

**Abstract**

Purpose: We present a new optimization technique for planning single arc VMAT (volumetric modulated arc therapy).

Methods: First, a convex multicriteria dose optimization problem is solved for an angular grid of 180 equi-spaced beams. This allows the planner to navigate the ideal dose distribution Pareto surface and select a plan of desired target coverage versus organ sparing compromise. The selected plan is then made VMAT deliverable by a simple fluence map merging and sequencing algorithm, which combines neighboring fluence maps based on a similarity score and then delivers the merged maps together, simplifying delivery. Successive merges are made as long the dose distribution quality is maintained.

Results: The method is applied to three cases: a prostate, a pancreas, and a brain. In each case, the Pareto selected plan is matched almost exactly with the VMAT merging routine, resulting in a high quality plan delivered with a single arc in less than three minutes on average.

Conclusions: The presented method offers significant improvements over existing VMAT algorithms. The first is the multicriteria planning aspect, which greatly speeds up planning time and allows the user to select the plan which represents the most desirable compromise between target coverage and organ at risk sparing. The second is the (user-chosen) epsilon-optimality guarantee of the final VMAT plan. Finally, the user can explore the tradeoff between delivery time and plan quality, which is a fundamental aspect of VMAT that cannot be easily investigated with current commercial planning systems.

## 1 Introduction

In the late 1990s, intensity modulated radiation therapy came on the clinical scene and quickly rose to a dominant position in radiation treatment. The relatively simple idea behind IMRT – to block flat radiation fields with the leaves of a collimator in order to produce spatially modulated fields – took time to realize due to both hardware and computational challenges. In 1982, Brahme described the dosimetric advantages of modulated beam fluence profiles [1]. Seven years later, both Webb and Bortfeld developed algorithms for optimizing fluence maps [2, 3], but it was not until seven years later that commercial IMRT systems (MLCs and their associated control systems) became available [4].

It is interesting to note that already by 1995, i.e. before commercial IMRT systems were available, rotational arc therapy, where the gantry rotates while MLC modulated beams are being delivered, was proposed [5]. It is only in the last couple of years however that hardware and software vendors have made rotational therapy commercially available (with the exception of TomoTherapy, which is a rotational therapy technique but is not considered in this paper because the optimization problem is fundamentally different), and this is largely due to the difficulty in treatment planning for such a large scale problem.

VMAT (volumetric modulated arc therapy, which is the term we will use throughout for rotational therapy delivered with a linac and an MLC) is a larger optimization problem than IMRT because it delivers radiation from every angle around the patient, and therefore dose computations need to be done for far more angles than IMRT. An even bigger hurdle presented by VMAT optimization is due to the coupling between adjacent angles: for efficient VMAT delivery, one should not move the MLC leaves more than necessary between neighboring angles. If VMAT is to be optimized with delivery time in mind, leaf positions need to be accounted for, which results in a large scale non-convex optimization. This non-convexity arises due to the non-linear mapping from leaf position to voxel dose: if one plots dose to a given voxel versus leaf position for a single leaf, the result is a sigmoid-shaped curve.

If delivery time would not be considered, then VMAT optimization is equivalent to a large IMRT optimization problem, and could therefore be solved by any of the methods developed for IMRT over the last 15 years. However, VMAT is about more efficient radiation delivery, and thus a VMAT optimization system should allow the user to select an appropriate compromise between delivery time and dose distribution quality.

The computational challenges of VMAT optimization have a direct impact on clinical VMAT usage. While the radiation therapy community generally agrees that VMAT plans are as good as or superior to IMRT plans, it is also well known that VMAT planning remains a great challenge and can be much more time consuming than IMRT planning [6, 7]. Multicriteria optimization (MCO) has been shown to be successful in reducing the planning time and increasing the plan quality for IMRT [8]. Since VMAT is a more challenging planning problem, MCO has the potential for even greater impact here.

It is possible to blindly write down the VMAT optimization problem and then apply various optimization algorithms to try to solve it. However, due to the complexity of the problem, we feel it is more useful, both for algorithm development and algorithm exposition, to first clearly understand the physical basis (hardware, treatment dose parameters, etc.) of the VMAT problem. To that end, we take the next couple of paragraphs to describe the relevant details.

Assuming a single fraction delivers 2 Gy to the target, if radiation were delivered to the patient via a single open field, it could be done in 200 monitor units (100 MU = 1 Gy is a standard MU calibration). A typical dose rate is 600 MU per minute, which means that it would take 20 seconds

to deliver the 2 Gy with a 10×10 cm field. At a maximum gantry speed of 6 degrees per second, a single revolution takes 60 seconds. Therefore, for a single revolution VMAT plan at maximal dose rate and gantry speed, one will see mostly small segments (on the rough order of $20/60 = 1/3$ the size of a 10×10 cm field). If one is willing to have the beam slow down to an average of half speed, completing the single arc in 144 seconds, one would see on average segments with a further 50% reduction in size.

With a maximum leaf speed of 2.5 cm/sec, leaves can travel across a 10 cm field in 4 seconds, in which time the gantry can rotate up to 20 degrees. To deliver highly modulated fields that are spaced close together, the gantry may need to slow down. In general, in a VMAT delivery the gantry needs to slow down when it takes longer to deliver the fluence pattern (required over some arc portion) than can be done at top gantry speed. It is useful to break this situation into two cases that represent the two causes for gantry slowing. The first case is when the fluence map has fluence levels that exceed the maximal fluence level that can be delivered at top gantry speed over the given arc portion. This clearly requires the gantry to slow down. The second case is when the field is modulated so much that it takes more time than available at top gantry speed to deliver all the isolated humps of the fluence map. The reason to distinguish these two cases is the following: in the first case, assuming the fluence profile is flat but at a large value, leaf speed is not the limiting factor: the gantry needs to slow down just to get enough dose in. For the second case however, the modulated fields might be able to be delivered without slowing down the gantry if the leaves could move fast enough.

A useful relationship here is that delivery time for a fluence map to be delivered by a left-to-right leaf sweep across the field is equal to the time it takes for the leaves to cross the field at top leaf speed plus the sum-of-positive-gradients (SPG) for the field (see Equation 3). The SPG for an IMRT field is a measure of the "ups and downs" of the field (the precise mathematical description is given in [9], and also briefly in the paragraph herein right before Equation 3). SPG can be minimized exactly in a convex optimization framework, whereas leaf travel distance, if incorporated up-front in the optimization, results in a non-convex problem. In our approach, we handle leaf travel issues in a VMAT-customized fluence map merging-and-sequencing routine which explicitly ensures that the dose distribution quality is maintained, while the delivery efficiency is successively improved. Our algorithm is designed to solve one of the key design issues of VMAT planning: where to optimally slow down the gantry. By merging like neighboring fluence maps and validating that the dose distribution after the merge is still good, we eliminate unnecessary gantry slow downs which arise from "over-delivery" of fluence maps. With our approach, the leaves travel back and forth at a high frequency only when needed and likewise the beam slows down only when necessitated by leaf travel requirements or SPG requirements.

VMAT treatments are currently delivered with Elekta [10] and Varian [11] equipment, and VMAT-like deliveries have been recently reported using Siemens equipment [12]. The treatment

delivery systems deployed by the different manufacturers have different designs and thus impose different delivery constraints in treatment planning. The Elekta and Varian linacs both allow dynamic machine parameter changes during the irradiation whereas for Siemens the delivery proceeds via a burst mode in a step and shoot fashion. For the Siemens system, where dynamic delivery restrictions do not play a role, a sequencer such as that advanced by [13] should be used to minimize the total beam on time in MU and the number of beam apertures required. For the dynamic VMAT deliveries the most important single constraint is the finite maximum MLC leaf velocity restriction. This limits the degree by which an aperture shape can change between two control points for a given dose rate. In our work, we consider only dynamic VMAT deliveries.

All dynamic arc planning algorithms approximate the continuous beam as a series of discrete static beams. Approaches to optimizing the final plans include both one and two step methods. In one step planning, the MLC motions are directly optimized with considerations for the limitations of the MLC and gantry motions, ensuring that the plans are able to be implemented on the linac. In two step planning, fluence maps are first optimized independently of delivery constraints. A leaf sequencing algorithm is then employed to convert the optimal fluence maps into deliverable MLC trajectories. A full review of VMAT optimization techniques is provided by Yu [14]. Here we briefly discuss two approaches which exemplify the main techniques used for VMAT planning.

In 2007, Varian adopted a one step algorithm for single-arc VMAT, reported by Otto [11], under the tradename RapidArc$^{\text{TM}}$. The method first optimizes the MLC motions for a coarse sampling of static points. Finer sampling is achieved by iteratively adding samples interpolated between existing static points until the desired sampling frequency is reached. Throughout, leaf positions are modified by local random search. Importantly, this algorithms allows both the gantry rate and dose rate to change along the arc.

Wang and Luan developed a two-step planning algorithm for single-arc VMAT that utilizes the graph theoretic concept of a shortest path to complete their leaf sequencing [15, 16]. Fluence maps spaced at 10° are first optimized using a conventional IMRT inverse planning algorithm. Leaf sequences are then determined by finding the shortest path on a directed acyclic graph consisting of all possible leaf positions for $k$ angles. The shortest path is the one that best minimizes the error, for a given delivery time, between the deliverable intensity profiles and the optimized fluence maps. A treatment time constraint is calculated before leaf sequencing, and reflects the number of arc portions to be sequenced and their required number of monitor units.

In this work, we provide a two step approach to VMAT planning that utilizes a multicriteria optimization algorithm to optimize 180 static beams placed at 2° intervals. Leaf sequencing is accomplished using a unidirectional sequencing algorithm. After obtaining this initial plan, neighboring fluence maps are iteratively merged to increase gantry speed and decrease delivery time. In this way, we work from the ideal solution towards one that is epsilon close to dose optimality, but has greatly increased delivery efficiency.

## 2 Methods

We begin by solving a 180 equi-spaced beam IMRT problem. We solve a multicriteria version of the IMRT optimization problem, which allows the planner to explore the tradeoffs between target coverage and healthy organ sparing, finally choosing a best-compromise solution [8, 17, 18]. Such a solution represents an ideal dosimetric plan, where treatment time is ignored. To actually deliver this solution, one would deliver the full IMRT fluence maps at every 2 degrees, which would be time consuming. Instead, we successively coarsen this 180-beam fluence map solution such that the delivery is made faster while the dose quality is kept within user selected bounds. Thus, in the sequencing step, we allow the user to explore the tradeoff between dose quality and delivery time.

In the following sections we describe the details of each of these components of our VMAT planning approach.

### 2.1 180-beam IMRT solution and Pareto surface plan selection

We consider the following multicriteria IMRT problem:

$$\begin{aligned} \text{optimize} \quad & \{g_1(d),\ g_2(d), \dots g_N(d)\} \\ \text{subject to} \quad & d = Df \\ & d \in C \\ & f \geq 0 \end{aligned} \tag{1}$$

Here $d$ is the vector of voxel doses, $D$ is the dose-influence matrix, and $f$ is a concatenation of all the fluence maps into a single beamlet fluence vector. The constraint set $C$ is a convex set of dose constraints. This can include for example bounds on mean structure doses, and minimum and maximum doses to individual voxels.

The objective functions are $g_1(d), \dots, g_N(d)$ where $N$ is the number of objectives defined. The optimization objectives can be any of the following: minimize the maximum structure dose, maximize the minimum structure dose, or minimize or maximize the mean structure dose. In general any convex functions would be permissible [19]. For our optimization, we only consider these ones since they can be handled with a linear solver, and since in the multicriteria planning context, they are typically sufficient to create high quality treatment plans [20, 21].

We solve this problem multiple times, approximating the Pareto surface, by following the methods detailed in [22]. Briefly, this method uses a feasibility projection solver that iteratively projects onto violated constraints until all constraints are satisfied. Objectives are turned into constraints with initially loose bounds which are gradually tightened until they are within user specified tolerance of optimality. After the projection solver runs for the $N$ objectives and some mixed objective

plans, the user navigates the solution space, which amounts to choosing the most preferable convex combination of the calculated Pareto surface plans. This plan, which we consider the ideal dosimetric plan, is then passed to the leaf sequencing and merging routine, described below.

### 2.2 Unidirectional leaf sequencing

In leaf sequencing the task is to create a set of MLC leaf trajectories which produce the desired fluence map while the gantry rotates over the arc portion allotted to that map. Each arc portion is assumed to be small, such that the angular difference in the ray paths produced by the rotating gantry is negligible as compared with the many static beam angles at which the fluence maps were optimized. To be deliverable, the leaf trajectories must not have leaf velocities greater than a given maximum value, either within the delivery of a given fluence map or between the delivery of one map and the next. A simple way of ensuring this condition is met is to sequence the trajectories as an alternating sequence of left to right and right to left dynamic MLC (dMLC) leaf sweeps. All leaves are aligned at one edge of the field at the beginning of the arc portion delivery and align at the opposite edge of the field at the end of the arc portion ready to commence the next arc portion with the leaves moving in the opposite direction.

The dMLC leaf sweep trajectory is calculated using the equations provided by [23, 24, 25], which give the leaf velocity of the leading ($v_{\text{lead}}$) and trailing ($v_{\text{trail}}$) leaves in terms of the maximum leaf velocity ($v_{\max}$) and the local fluence gradient $\frac{\mathrm{d}f(x)}{\mathrm{d}x}$. The equations (2) give the leaf velocities in terms of bixels per MU delivered and require a constant dose rate over the arc portion.

$$\begin{aligned} &\left( v_{\text{lead}}(x) = v_{\max},\ v_{\text{trail}}(x) = \frac{v_{\max}}{1 + v_{\max}\frac{\mathrm{d}f(x)}{\mathrm{d}x}} \right) \quad \text{if} \quad \tfrac{\mathrm{d}f(x)}{\mathrm{d}x} \geq 0 \\ &\left( v_{\text{trail}}(x) = v_{\max},\ v_{\text{lead}}(x) = \frac{v_{\max}}{1 - v_{\max}\frac{\mathrm{d}f(x)}{\mathrm{d}x}} \right) \quad \text{otherwise} \end{aligned} \tag{2}$$

The time for all leaf pairs to traverse the field is given by (3) and is governed by the width of the field $W_F$ and the ratio of the maximum over the sum of positive gradients terms ($\sum \frac{\mathrm{d}f(x)}{\mathrm{d}x}^{+}$) as evaluated over each leaf path divided by the dose rate $r$.

$$T = \frac{W_F}{v_{\max}} + \frac{\max_{\text{rows}}\left(\sum \frac{\mathrm{d}f(x)}{\mathrm{d}x}^{+}\right)}{r} \tag{3}$$

Each fluence map is locked to a given portion of the gantry rotation arc, so that if the gantry rotation time over the arc portion is less than the leaf travel time required, the gantry speed is reduced. In a similar manner, if the required leaf travel time is less than the gantry rotation time, then the leaf travel time is increased by reducing the dose rate over the arc portion. It should be noted that currently a continuously variable dose rate is assumed, but also that there is no strict requirement that the leaves take the full duration of the time available to complete the

fluence modulation. However, it is dosimetrically favorable to reduce the dose rate if this can be accomplished without an effect on delivery time, as this will lead to larger beam apertures (all else being equal, larger aperatures are preferred due to reduced scatter and higher confidence in the associated dose calculations).

The leaf velocities of the leading and trailing leaves are then assigned to the right and left leaves respectively. The next fluence map is processed in the opposite direction (right to left) and the leading and trailing leaf trajectories assigned to the left and right leaves respectively. This process then repeats for all the fluence maps in the VMAT arc. The fluence produced by the final delivery control points is then computed over the original 2° angular bins to facilitate the fluence map merging process.

### 2.3 Merging neighboring fluence maps

The purpose of the merging algorithm is to lower the beam-on treatment time by reducing the number of distinct fluence maps that need to be delivered. To deliver each fluence map, the leaves must make a full unidirectional sweep across the aperture over the arc portion that fluence map is specified for. VMAT solutions with a large number of distinct fluence maps thus require the gantry to move slowly in order to give the leaves sufficient time to move across the field. Our merging algorithm iteratively merges neighboring fluence maps, allowing the gantry to move more quickly around the full-arc.

We begin with 180 optimized fluence maps which are delivered over the ranges $[0, 2°], [2°, 4°]$, ...$[358°, 360°]$. The initial solution is a high-quality treatment plan, and we seek to merge fluence maps in a way that preserves this optimized dose distribution. Our merging strategy is based on the observations that 1) merging fluence maps with the greatest degree of similarity will have the least effect on the final dose distribution, and 2) merging fluence maps with small arc portions will have less of an effect on the dose distribution than merging fluence maps defined over long arc portions.

These two observations allow us to define a similarity metric between any two neighboring fluence maps $f^1$ and $f^2$, with arc portion lengths of $\theta_1$ and $\theta_2$. The similarity metric $\delta$ is defined as the Frobenius norm of the difference between the maps (normalized by their arc portion lengths to make them comparable), scaled to the combined arc portion length $\theta_1 + \theta_2$:

$$\delta(f^1, f^2) \quad = \quad (\theta_1 + \theta_2) \sqrt{\sum_{i,j} \left( \frac{f^1_{ij}}{\theta_1} - \frac{{f^2}_{ij}}{\theta_2} \right)^2}. \tag{4}$$

We incorporate this similarity metric into a greedy search algorithm that merges a single pair of fluence maps with every iteration, such that after $n$ iterations the number of fluence maps is $180 - n$. The neighboring pair with the lowest $\delta$ score is selected for merging. The merged fluence

map is defined as the sum of the two neighboring fluence maps, with a new arc portion equal to the union of the initial two arc portions. This combined map is then sequenced and the fluence is binned into the original 2° bins. The stopping criterion for the greedy search will depend on the planner's desired balance between treatment time and plan quality.

### 2.4 Sensitivity to algorithm settings

We apply the technique to three different clinical cases. For the prostate case, we also investigate some variations to the algorithm. The first is smoothing the solution after the Pareto navigation phase. We consider two types of smoothing. The first smoothing method minimizes the maximum beamlet value with all the objective values of the ideal dosimetric solution turned into constraints, also maintaining the original MCO formulation constraints. The second smoothing method uses an SPG smoother during the solver's projection steps. During the projection iterations, an SPG smoothing step is periodically called. This step identifies the single row with the largest SPG and redistributes the fluences by reducing the peak fluences of that row by a factor of 0.9, and then adding the 10% to the neighboring adjacent beamlets. This is a heuristic approach to controlling the SPG with a projection solver inspired by smoothing kernels in projection-based image reconstruction [26].

We also investigate how the final VMAT solution depends on beamlet size and beam angle spacing, in order to show that our solution technique yields a fundamentally correct VMAT plan and not one that is sensitive to algorithm initial conditions. Since most commercial VMAT solutions calculate the final dose on a 2 degree gantry spacing, we choose to use this as our baseline angular spacing grid. We examine the nature of the VMAT solution that arises when this grid is coarsened to 4 degrees (thus, we start by solving a 90 beam IMRT problem). We then take this angular grid and further investigate shrinking the beamlet size by a factor of 2 in the leaf travel direction (creating 0.5×1 cm beamlets).

## 3 Results

We demonstrate the method on three clinical cases: a prostate, a pancreas, and a brain case with two distinct targets. For each case we assume we are designing a 2 Gy fraction plan. It is important to note that unlike step and shoot IMRT optimization, where fraction dose scaling does not fundamentally affect the plan, here fraction dose is important since it is linked to dose rate, gantry speed, and leaf speed (see Discussion). For display purposes we scale the dose-volume-histograms (DVHs) up to the total dose delivered from all the fractions. We also display the optimization formulations for the total dose. We use CERR 3.0 beta 3 [27] for dose computation. We use the following VMAT delivery parameters: maximum gantry speed = 1 rotation/min, maximum leaf speed = 2.5 cm/sec, maximum dose rate = 600 MU/min.

For the prostate case (voxel size: 3×3×2.5mm, 1×1cm beamlets), the main dosimetric tradeoff is between the rectum dose and the target coverage. However, for this analysis we hold the prostate coverage fixed at the prescription level, 79 Gy, and consider the tradeoff between mean dose to the rectum (more precisely, the anterior rectum as contoured by the physician), the bladder, and the unclassified tissue. A Pareto surface is constructed using the following multicriteria formulation:

$$\begin{aligned} \text{minimize} \quad & \{\text{mean rectum dose, mean bladder dose, mean u.t. dose}\} \\ \text{subject to} \quad & d = Df \\ & d_i \geq 79 \text{ Gy}, \ \forall i \in \text{ target} \\ & d_i \leq 45 \text{ Gy}, \ \forall i \in \text{ femoral heads} \\ & d_i \leq 79 * 1.07 \text{ Gy}, \ \forall i \\ & f \geq 0 \end{aligned} \tag{5}$$

where u.t. stands for unclassified tissue: all the voxels not belonging to any other structure.

We navigate to a solution with (mean rectum dose, mean bladder dose, mean u.t.)=(0.50, 0.50, 0.12)*79 Gy. The results of the sequential merging routine are shown in Figure 1. Plan quality is assessed using the mean dose to the anterior rectum, femoral heads, and bladder, the standard error from the prescription level of the dose to the prostate, and the volume of the prostate receiving the prescription dose.

Figure 1 shows the results of the merging algorithm for the prostate case. The plan does not degrade until after 140 iterations, where we begin to lose target coverage. Therefore, we selected a plan requiring 40 arc portions and a treatment time of 187.4 s (original time of 806.2 s with 180 arc portions), which had the optimal tradeoff between treatment time and plan quality. This tradeoff between quality and time is shown for both the target (Figure 1a) and three key organs-at-risk (Figure 1b). The DVH comparing the original plan (solid line) to the simplified plan (dashed line) is shown in Figure 1c. The simplified plan has (mean rectum dose, mean bladder dose)=(0.50, 0.50)*79 Gy, with 98.9% of the prostate volume receiving the full dose (V79Gy). Femoral head constraints are easily maintained. The plot of the arc portions in Figure 1d shows that the gantry speed varies during the course of the single arc. Smaller arc portions require the gantry to slow down to allow the leaves time to traverse the aperture, and represent the fluence maps that are most dissimilar to their neighbors.

For the prostate case, we additionally examine the sensitivity of our approach to the following issues: smoothing of the IMRT solution before passing it to the merging routine, using fewer than 180 beams, and using smaller beamlets. The purpose of this experiment is to verify that quality of the final plan remains generally insensitive to the initial smoothing method, and that the use of 1 cm beamlets does not bias our results in any way.

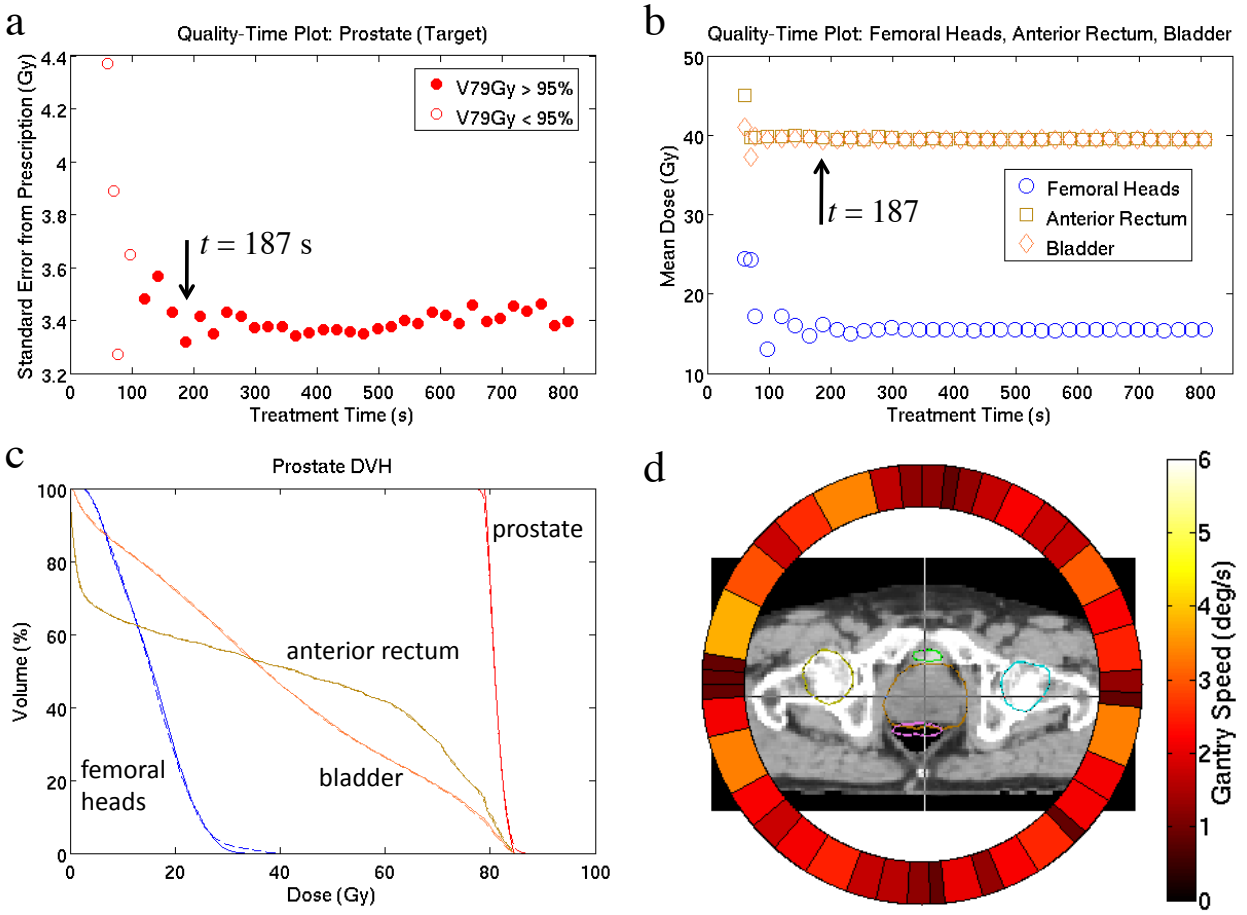


Figure 1: Results of the merge algorithm for prostate VMAT. The merged plan that was determined to have the best tradeoff between quality and treatment time is indicated by the black arrow. Quality-time tradeoff plots are shown for the a) prostate target and b) the anterior rectum. c) The DVH data for the original (solid) and merged (dashed) plan. d) The fluence map arc portion plot for the merged plan, showing the gantry speed at different angles.

The total SPG values (in seconds) for no smoothing, max beamlet smoothed, and SPG smoothed, are 168, 113, and 92.2 respectively. Figure 2a shows the DVH plots for the final merged plans for these three smoothing methods. Each plan is simplified from an initial 180 beam solution, and represents the point on the tradeoff curve with the best compromise between plan quality and treatment time. The three plans are highly similar, differing only slightly in the dose to the femoral heads, which is simply reflective of the original plan. Because the type of smoothing does not have a significant effect on the quality-time tradeoff, we will use only max beamlet smoothing for our three disease sites.

Figure 2b shows DVH plots for final plans that were created by iterative merging from an initial 180 beam solution (1×1 cm beamlet), 90 beam solution (1×1 cm), and 90 beam solution with reduced beamlet sizes (.5×1 cm). The final plans have similar treatment times and target

coverage. In the optimization with the smaller beamlets, we reduce the rectum dose down as much as possible. As expected, finer beamlet resolution yields better dose distribution shaping (the mean anterior rectum dose is reduced from 50% to 44% of the prescription dose), which indicates the value of reducing the beamlet size. But to ease the computational burden we remain with the 1 cm beamlets for the pancreas case. We switch to .5×.5 cm beamlets for the brain case due to its overall smaller target.

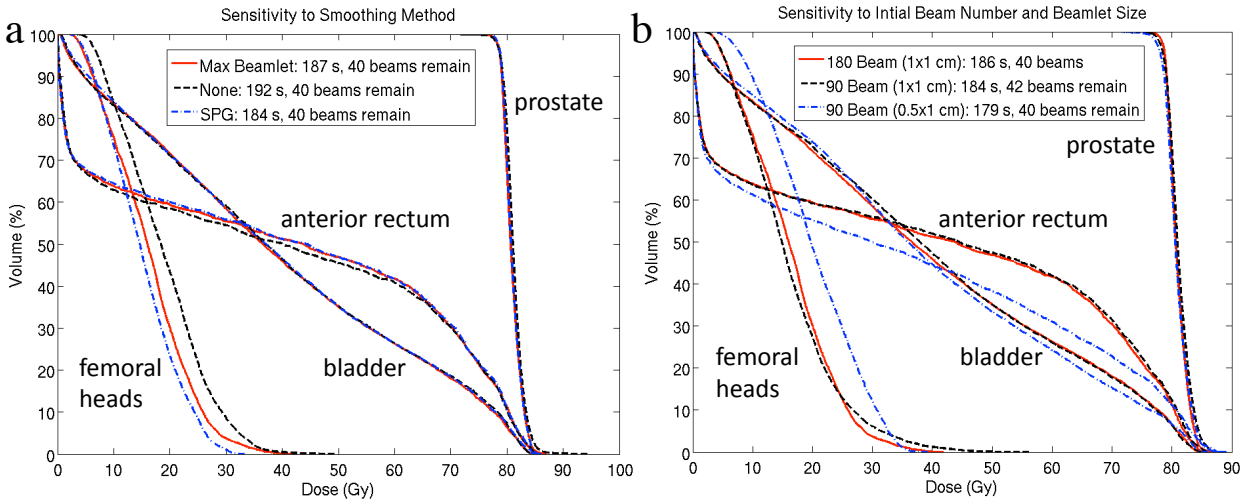


Figure 2: The sensitivity of our algorithm to different initial plans was tested on our prostate case. a) DVH plots for the final merged plans for 3 different 180 beam initial plans: max beamlet smoothing, no smoothing, and SPG smoothing. b) DVH plots for the final merged plans for 3 different initial plans, all with max beamlet smoothing: 180 beams, 90 beams, and 90 beams with small beamlets (.5×1 cm).

The second case is a pancreas case (voxel size: 2.6×2.6×2.5 mm, 1×1 cm beamlets). The pancreas is an interesting site for VMAT for the same reason that it is interesting for beam angle optimization: it is surrounded by kidneys, stomach, and liver, and the optimal radiation entry directions are not obvious [28]. Our MCO formulation is as follows:

$$\begin{aligned} \text{minimize} \quad & \{\text{mean shell dose, mean kidneys dose, mean liver dose,} \\ & \text{mean stomach dose, -min target dose}\} \\ \text{subject to} \quad & d = Df \\ & d_i \geq 50.4 * 0.95 \text{ Gy}, \ \forall i \in \text{ target} \\ & d_i \leq 45 \text{ Gy}, \ \forall i \in \text{ spinal cord} \\ & d_i \leq 50.4 * 1.12 \text{ Gy}, \ \forall i \\ & f \geq 0 \end{aligned} \tag{6}$$

The shell is a .7 cm band around the target used to promote dose conformality to the target. We

navigate to a solution with (mean shell dose, mean kidneys dose, mean liver dose, mean stomach dose)=(.9, .24, .23, .20, .95)*50.4 Gy. The maximum beamlet level can then be reduced to 0.4 (and possibly beyond, but this is the value of the maximum level that can be delivered in 2 degrees without slowing the gantry down, so we do not attempt to reduce it further).

Figure 3 shows the results of the merging algorithm for the pancreas case. Plan quality begins to degrade after 155 merges, after which the mean dose to the kidneys, liver and stomach begins to increase. The selected plan requires 25 fluence maps and 108.4 s (initial time of 732.3 s), and reflects nearly the same DVH as the original 180 beam solution. The selected plan has (mean kidneys dose, mean liver dose, mean stomach dose)=(.24, .23, .21)*50.4 Gy, with 95.8% of the tumor volume receiving the full dose (V50.4Gy).

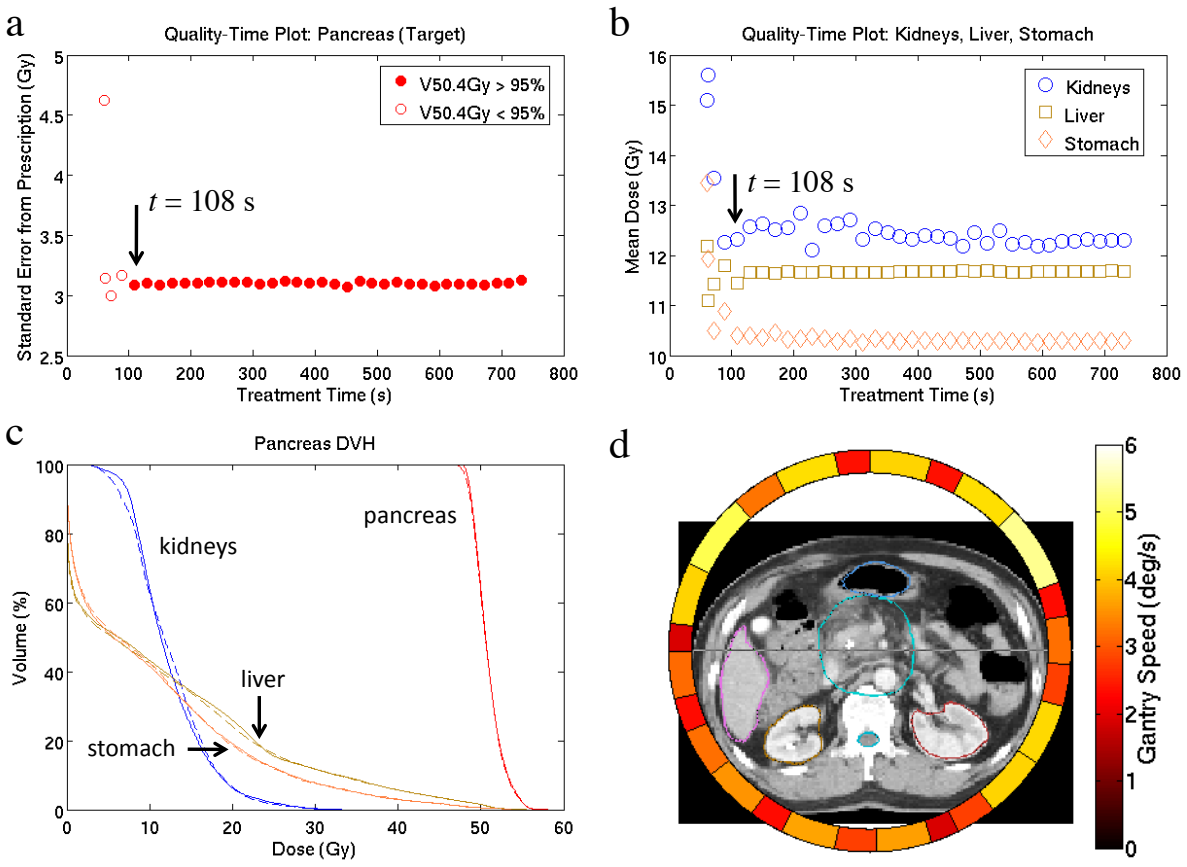


Figure 3: Results of the merge algorithm for pancreas VMAT. The merged plan that was determined to have the best tradeoff between quality and treatment time is indicated by the black arrow. Quality-time tradeoff plots are shown for the a) pancreas target and b) kidneys, liver and stomach. c) The DVH data for the original (solid) and merged (dashed) plan. d) The arc portion plot for the merged plan, showing the gantry speed at different angles.

The third case is a brain double lesion case (voxel size: 1.35×1.35×1.25 mm, .5×.5 cm beamlets). We selected this case due to the potential of VMAT to treat isolated metastic lesions in a single

gantry revolution. Minimizing treatment time in these cases is important because of the precise set up required for such treatments. In this case there are two adjacent lesions, one of them being inside the brainstem, and the prescription is 32 Gy. For the optimization, we consider the targets as a single combined structure:

$$\begin{aligned} \text{minimize} \quad & \{\text{mean chiasm dose, mean brainstem dose, mean u.t. dose, -min target dose}\} \\ \text{subject to} \quad & d = Df \\ & d_i \geq 32 * 0.95 \text{ Gy}, \ \forall i \in \text{ target} \\ & d_i \leq 32 * 1.15 \text{ Gy}, \ \forall i \\ & f \geq 0 \end{aligned} \tag{7}$$

We navigate to a solution with (mean chiasm dose, mean brainstem dose, mean u.t. dose)=(.15, .9, .44)*32 Gy. The maximum beamlet level is reduced to 0.9 and then the solution is passed to the merging algorithm.

Figure 4 shows the results of the merging algorithm for the brain case. Plan quality begins to degrade after 145 merges, after which the mean dose to the optic chiasm begins to rise. The selected plan requires 35 fluence maps and 138.6 s (initial time of 662.0 s), and reflects nearly the same DVH as the original 180 beam solution. The selected plan is associated with (mean chiasm dose, mean brainstem dose=(.15, .9)*32 Gy, with 99.0% of the target volume receiving the full dose (V32Gy).

## 4 Discussion and Conclusions

Finite leaf speed is the single parameter that makes the single arc coplanar VMAT optimization problem so challenging. As leaf speed approaches infinity, the complexity and delivery time of a plan is governed solely by the SPG of the fluence maps, and this quantity can be minimized exactly in a convex optimization setting [9]. On the other hand, for fields even with low SPG in the finite leaf speed setting, the beam may have to slow down just to have the leaves travel across the field. This information cannot be represented in a convex optimization setting. Fortunately, similar to the decomposition of IMRT planning into a convex fluence map optimization step and then a leaf sequencing step, we show in this work that we can similarly decompose the VMAT problem. The additional challenge in the VMAT setting stems from the continous motion of the leaves and the gantry.

We adopt an approach of starting with a fine solution and gradually coarsening it. This decreases the delivery time while maintaining a good dose distribution. The rationale for our fine-to-coarse approach is that in VMAT planning, whatever method is used to derive a solution, one will ultimately do a dose computation on a fine angular grid, such as a 2 degree spacing. Given current computing capacity and tailored algorithms to solve the IMRT problem (e.g. [22, 29]), it is not

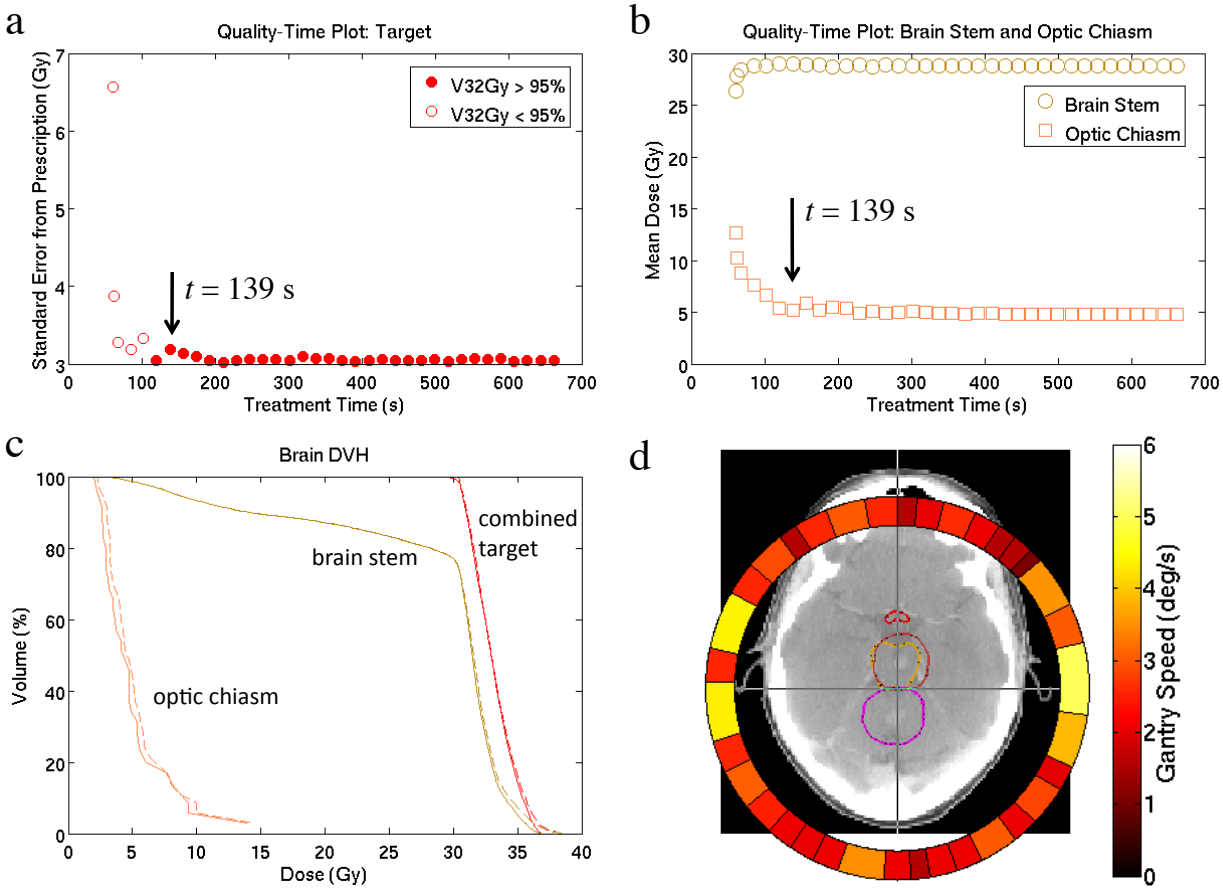


Figure 4: Results of the merge algorithm for brain VMAT. The merged plan that was determined to have the best tradeoff between quality and treatment time is indicated by the black arrow. Quality-time tradeoff plots are shown for the a) tumor target and b) the brain stem and optic chiasm. c) The DVH data for the original (solid) and merged (dashed) plan. d) The fluence map range plot for the merged plan, showing the gantry speed at different angles.

difficult to solve a 180-beam IMRT optimization problem. We show in this paper that starting from this ideal plan, we can successively smoothen the delivery until we have a plan that is both of high dose quality and efficient to deliver. The user first explores the dosimetric tradeoffs exactly as done for IMRT MCO, and then either presses a button which does automatic merging to a pre-specified epsilon dose deviation threshold, or uses the merging routine to interactively explore the tradeoff between dose quality and delivery time, i.e. the information displayed in Figures 1, 3, and 4. To create these figures, we run the merging algorithm until the gantry is moving at maximum speed for the entire rotation, in order to show the entire plan quality-delivery time tradeoff curve. This merging takes on the order of 5 minutes for each of the cases (this includes computing the dose and plan evaluation metrics after each merge; if you knew a priori how many merges you wanted, the merging would take on the order of 10 seconds). Our method of starting from the fine ideal

solution and then making it deliverable contrasts with all of the other VMAT approaches which start from a coarse solution and then add segments to improve the dose quality [11, 14]. Another way to view the difference is: we start at the global minimum of a relaxed convex problem and then modify that solution to make it deliverable while staying as close as desired to the ideal plan, while the coarse-to-fine approaches first employ a global search and then use local refinements to optimize further but do not yield any guarantees on optimality. Such guarantees are important both for high quality patient care and for proper comparisons of treatment modalities, comparisons which can easily be muddled when using planning systems with no optimality guarantees.

Our VMAT approach is designed to intrinsically address two closely related VMAT issues: 1) how much fluence modulation is needed as the gantry proceeds around the patient, and 2) with what frequency do the fluence maps need to be delivered. The second point is an important one in the context of minimizing delivery time in the finite leaf speed setting. Using a finely spaced angular grid of beams and an IMRT solver to compute optimal fluence maps, one may observe that neighboring maps are identical. Indeed, in symmetric cases like a donut-shaped target with a central circular organ at risk, the maps from each angle are identical, and yet one does not know from this how frequently those modulated fields need to be formed by the MLC leaves as the gantry rotates around the patient. In the merging algorithm, we not only combine fluence maps, but we also enforce that the combined map will be delivered once over the span occupied by the two original fluence maps. When the user makes the final delivery selection, leaf motion throughout the gantry rotation occurs at a rate dependent on how many merges occured. Even with our simple greedy selection strategy (merging the current two neighboring fluence maps with the best similarity score and employing with no backtracking), we see that we can merge many of the maps and not change the plan perceptibly. It remains to be investigated if something other than the greedy merging strategy, and/or merging based on dose distribution similarities rather than fluence map similarities, could be more successful.

We study two approaches to create initially smoother 180-beam IMRT fluence maps. Neither of the two methods has a dramatic effect on the final dose distribution or treatment time, although they both show improvements in the expected direction. Max beamlet smoothing reduces treatment time by 5 seconds (a 2.6% decrease from no smoothing), and SPG smoothing by another 3 seconds. Because the merging routine combines neighboring fluence maps, any noise in the maps that is not dosimetrically meaningful or useful will tend to get washed out. We speculate that this is why we do not observe large gains from initial fluence map smoothing. For practical implementation, given the difficulty of implementing an SPG smoothing solver, we recommend either no smoothing or max beamlet smoothing.

Although we used fairly simple formulations for the initial 180-beam IMRT solution, the method proposed herein does not preclude the use of more sophisticated formulations. For example, quadratic penalty formulations, dose-volume constraints, equivalent uniform dose, and biologi-

cal objective functions could all be used. Also, there is flexibility in the sequencing/merging step, and any number and type of user-defined cutoff values could be defined to determine what solution along the delivery time/plan quality curve should be delivered. We have also used the simple pencil beam decomposition approach for the IMRT and VMAT problem. The dose distribution produced by such "Dij" approaches, in naïve implementations, can degrade significantly when final dose calculations – which include output factor corrections and leaf transmissions – are performed, but we have proceeded with the Dij approach due to its mathematical tractability and because it is possible to successfully include delivery effects into Dij-based approaches [30].

We have considered unidirectional leaf sequencing in this paper. It is possible that more general leaf sequencing would allow one to deliver the merged maps more quickly, but for fluence rows with multiple ups-and-downs, one needs to sweep both leaves across almost the entire row to achieve the inner row modulations. We do not believe there will be dramatic gains in switching to more general leaf sequencing, but we do plan on investigating this.

Other extensions to the proposed approach include collimator rotation, couch rotation, and dual and partial arc considerations. There could be a reduction in treatment time by aligning the collimator to the fluence maps in such a way to minimize fluence map delivery time (Equation 3). The collimator angle trajectory could be determined by examining the fluence maps of the 180-beam IMRT solution with this metric, and therefore seems not an overly difficult extension of the approach, aside from the potentially irksome issue of forming fluence maps with a rotating MLC Couch rotation on the other hand, which leads to non-coplanar arcs, leads to a much larger optimization space and is accordingly a much more challenging problem.

A partial arc solution could save treatment time by allowing the gantry to spend more time at angles that are more beneficial for radiation delivery. A partial arc solution could be formed by prescribing an arc based on user experience, or by observing the original 180-beam solution and deciding to eliminate an angle sector which has generally low fluence. Once an arc is chosen, the multicriteria dose optimization and subsequent merging steps would proceed exactly as specified in this report.

Because we allow the gantry to slow down as much as possible to deliver the required fluence patterns, there can be no dosimetric advantages to double arc (or more) solutions. The only possible advantage for multiple arcs is treatment time. Double arc solutions will be treatment time superior when most of the fluence maps are highly modulated large fields which can be delivered quicker in two sweeps, with the leaves reset on the second sweep to be in favorable positions. For example, a double hump fluence map, with the humps separated by a wide zero fluence section, would be faster with a two arc approach if the first arc delivered the first hump and the leaves could be positioned correctly during the second arc to deliver the second hump. To make a double arc plan overall faster in treatment delivery, one would need a large number of the fluence maps to be of this nature. We speculate that such situations will not arise often clinically, and that therefore single

arc solutions, with good optimizers, will typically be the right choice. The single arc versus dual arc issue will be examined in a future publication.

The ratio of fraction dose to the treatment machine's maximum dose rate is an important quantity in VMAT planning. For hypofractionation, the fraction dose can be much higher than 2 Gy. This gives the delivery system more time for delivery and shifts the treatment time bottleneck from leaf speed to dose rate: since it will now take longer to deliver the required dose, the beam can slow down more overall and finite leaf speed becomes less influential. On the other hand, higher dose rates, achieved for example by treating without the flattening filter, will shift the burden to the finite MCL leaf speed again. These are important issues that should be kept in mind by VMAT researchers.

While interesting algorithmic challenges remain for VMAT (non coplanar arcs, dynamic collimator rotations, optimal partial and multiple arc creation), we have introduced a method for single arc coplanar VMAT that guarantees delivery of an epsilon-optimal dose distribution. No IMRT researcher would claim that one ever needed more than 180 equi-spaced beams for an optimal coplanar IMRT solution. Since we start with such a plan, and then make it VMAT-deliverable, our method guarantees a proveably optimal (up to an arbitrarily small user-specified tolerance) treatment plan, which is something no commercial VMAT planning system currently does.

**Acknowledgements** The project described was supported by Award Number R01CA103904 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. Thanks to Brian Winey for providing the brain case.

## References

[1] A. Brahme, J.E. Roos, and I. Lax. Solution of an integral equation in rotation therapy. *Physics in Medicine and Biology*, 27:1221–1229, 1982.

[2] S. Webb. Optimisation of conformal radiotherapy dose distributions by simulated annealing. *Physics in Medicine and Biology*, 34(10):1349–1370, 1989.

[3] T. Bortfeld, J. Bürkelbach, R. Boesecke, and W. Schlegel. Methods of image reconstruction from projections applied to conformation radiotherapy. *Physics in Medicine and Biology*, 35:1423–1434, 1990.

[4] S. Webb. The physical basis of IMRT and inverse planning. *British Journal of Radiology*, 76:678–689, 2003.

[5] C. Yu. Intensity-modulated arc therapy with dynamic multileaf collimation: an alternative to tomotherapy. *Physics in Medicine and Biology*, 40(9):1435–1449, 1995.

[6] M. Oliver, W. Ansbacher, and W. Beckham. Comparing planning time, delivery time and plan quality for IMRT, rapidarc and tomotherapy. *Journal of Applied Clinical Medical Physics*, 10(4):117–131, 2009.

[7] M. Rao, W. Yang, F. Chen, K. Sheng, J. Ye, V. Mehta, D. Shepard, and D. Cao. Comparison of Elekta VMAT with helical tomotherapy and fixed field IMRT: plan quality, delivery efficiency and accuracy. *Medical Physics*, 37(3):1350–1359, 2010.

[8] D. Craft, T. Hong, H. Shih, and T. Bortfeld. Improved planning time and plan quality through multicriteria optimization for intensity-modulated radiotherapy. *Int. J. Radiation Oncology Biol. Phys.*, 80(3):790–799, 2011.

[9] D. Craft, P. Süss, and T. Bortfeld. The tradeoff between treatment plan quality and required number of monitor units in IMRT. *Int. J. Radiation Oncology Biol. Phys.*, 67(5):1596–1605, 2007.

[10] J. Bedford and Warrington A. Commissioning of volumetric modulated arc therapy (VMAT). *Int. J. Radiation Oncology Biol. Phys.*, 73:537–545, 2008.

[11] K. Otto. Volumetric modulated arc therapy: IMRT in a single gantry arc. *Medical Physics*, 35(1):310–317, 2008.

[12] B. Salter, V. Sarkar, B. Wang, H. Shukla, M. Szegedi, and P. Rassiah-Szegedi. Rotational IMRT delivery using a digital linear accelerator in very high dose rate 'burst mode'. *Physics in Medicine and Biology*, 56(7):1931, 2011.

[13] K. Engel. A new algorithm for optimal multileaf collimator field segmentation. *Discrete Applied Mathematics*, 152(1-3):35 – 51, 2005.

[14] C. Yu and G. Tang. Intensity-modulated arc therapy: principles, technologies and clinical implementation. *Physics in Medicine and Biology*, 56(5):R31–R54, 2011.

[15] C. Wang, S. Luan, G. Tang, D. Chen, M. Earl, and C. Yu. Arc-modulated radiation therapy (AMRT): a single-arc form of intensity-modulated arc therapy. *Physics in Medicine and Biology*, 53(22), 2008.

[16] S. Luan, C. Wang, D. Cao, D. Chen, D. Shepard, and C. Yu. Leaf-sequencing for intensity-modulated arc therapy using graph algorithms. *Medical Physics*, 35(1):61–69, 2008.

[17] M. Monz, K-H. Küfer, T. Bortfeld, and C. Thieke. Pareto navigation - algorithmic foundation of interactive multi-criteria IMRT planning. *Physics in Medicine and Biology*, 53(4):985–998, 2008.

[18] C Thieke, K-H. Küfer, M Mon, A Scherrer, F Alonso, U Oelfke, P Huber, J Debus, and T Bortfeld. A new concept for interactive radiotherapy planning with multicriteria optimization: First clinical evaluation. *Radiotherapy and Oncology*, 85(2):292–298, 2007.

[19] H. Romeijn, J. Dempsey, and J. Li. A unifying framework for multi-criteria fluence map optimization models. *Physics in Medicine and Biology*, 49:1991–2013, 2004.

[20] D. Craft, T. Halabi, and T. Bortfeld. Exploration of tradeoffs in intensity-modulated radiotherapy. *Physics in Medicine and Biology*, 50(24):5857–5868, 2005.

[21] T Hong, D Craft, F Carlsson, and T Bortfeld. Multicriteria optimization in intensity-modulated radiation therapy treatment planning for locally advanced cancer of the pancreatic head. *Int. J. Radiation Oncology Biol. Phys.*, 72(4):1208–1214, 2008.

[22] W. Chen, D. Craft, T. Madden, K. Zhang, H. Kooy, and G. Herman. A fast optimization algorithm for multi-criteria intensity modulated proton therapy planning. *Medical Physics*, 37(9):4938–4935, 2010.

[23] R. Svensson, P. Källman, and A. Brahme. An analytical solution for the dynamic control of multileaf collimators. *Physics in Medicine and Biology*, 39:37–61, 1994.

[24] J. Stein, T. Bortfeld, B. Dörschel, and W. Schlegel. Dynamic x-ray compensation for conformal radiotherapy by means of multi-leaf collimation. *Radiotherapy and Oncology*, 32:163–173, 1994.

[25] S.V. Spirou and C.S. Chui. Generation of arbitrary intensity profiles by dynamic jaws or multileaf collimators. *Medical Physics*, 21(7):1031–1041, 1994.

[26] G. Herman. *Fundamentals of Computerized Tomography: Image Reconstruction from Projections*. Springer-Verlag, London, 2009.

[27] J. Deasy, A. Blanco, and V. Clark. Cerr: A computational environment for radiotherapy research. *Medical Physics*, 30(5):979–985, 2003.

[28] E. Woudstra. *Beam orientation selection in radiotherapy treatment planning*. PhD thesis, Technical University Delft, 2006.

[29] C. Men, X. Gu, D. Choi, A. Majumdar, Z. Zheng, K. Mueller, and S. Jiang. Gpu-based ultrafast IMRT plan optimization. *Physics in Medicine and Biology*, 54(21):6565–6573, 2009.

[30] U. Jelen, M. Sohn, and M. Alber. A finite size pencil beam for IMRT dose optimization. *Physics in Medicine and Biology*, 50(8):1747–1766, 2005.