Annals of Operations Research 119, 165–181, 2003

Integer Programming Applied to Intensity-Modulated Radiation Therapy Treatment Planning *

EVA K. LEE **
eva.lee@isye.gatech.edu
Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA and
Radiation Oncology, Emory University School of Medicine, Atlanta, GA, USA

TIM FOX and IAN CROCKER
Radiation Oncology, Emory University School of Medicine, Atlanta, GA, USA

Abstract. In intensity-modulated radiation therapy (IMRT) not only is the shape of the beam controlled, but combinations of open and closed multileaf collimators modulate the intensity as well. In this paper, we offer a mixed integer programming approach which allows optimization over beamlet fluence weights as well as beam and couch angles. Computational strategies, including a constraint and column generator, a specialized set-based branching scheme, a geometric heuristic procedure, and the use of disjunctive cuts, are described. Our algorithmic design thus far has been motivated by clinical cases. Numerical tests on real patient cases reveal that good treatment plans are returned within 30 minutes. The MIP plans consistently provide superior tumor coverage and conformity, as well as dose homogeneity within the tumor region while maintaining a low irradiation to important critical and normal tissues.

Keywords: intensity-modulated radiation therapy, external beam radiotherapy, optimization, mixed integer programming, treatment planning

AMS subject classification: 90C11, 90C90, 90-08

1. Introduction

Intensity-modulated radiation therapy (IMRT) is an important recent advance in radiation therapy. In conventional radiotherapy treatment, the planning process consists of determining a set of external beams that meet, as best as possible, the clinical dose distribution criteria. In many cases, significant compromises to critical structure function have to be made to enable a tumoricidal dose to be delivered to the targets. In IMRT, the radiation fluence is varied across the beam, which allows a higher degree of conformation to the tumor than previously possible and allows concave isodose profiles to be generated. Specifically, not only is the shape of the beam controlled, but combinations of open and closed multileaf collimators modulate the intensity as well. For this reason, IMRT provides improved delivery power over conventional treatment. Indeed, it provides an unprecedented capability to dynamically vary the dose to accommodate the shape of the tumor from different angles, and to spare normal tissues and organs-

* Results of this paper was presented at INFORMS Hawaii, June 2001.
** Corresponding author.

at-risk (OAR) that may be potentially harmed during treatment. However, due to the complexity of the beam intensity profile associated with IMRT, a computer-driven optimization algorithm must be used to determine the beam fluences (intensity maps) that provide the best compromise between target underdosing, target overdosing and critical structure overdosing. In this paper, we offer a mixed integer programming approach for determining an optimal configuration of intensity maps and beam angles for IMRT.

In section 2 we describe the treatment planning problem for IMRT, discuss relevant issues on dose calculation, and specify several mixed integer programming treatment planning models. Computational strategies for plan optimization are presented in section 3. Specifically, we discuss the implementation of a constraint and column generator, a specialized set-based branching scheme, a geometric heuristic procedure, and the use of disjunctive cuts. Our algorithmic design thus far has been motivated by clinical cases. Numerical results are analyzed in section 4. In section 5 we present some clinical interpretation of the solutions obtained via the MIP approach. This is followed by concluding remarks in section 6.

2. Intensity-modulated radiation therapy and treatment planning optimization

2.1. Background

Linear accelerators (LINACs) are common beam delivery units used for external beam radiotherapy. The table on which the patient lies and the beam delivery mechanism for the LINAC rotate about separate axes, providing the ability to adjust the angle and entry point of radiation fields used during treatments. Each field is further defined by a bank of multileaf collimators (MLC), small metallic leaves located inside the treatment unit (LINAC). These leaves can be opened or closed, and used to shape the radiation beam as it exits the machine. Figure 1 shows a linear accelerator.

For intensity-modulated radiation therapy (IMRT) [8,10,15,20,21,35,37], the shape of the beams, and the combinations of open and closed MLC leaves control and modulate the intensity. This provides the ability to dynamically vary the dose to accommodate the shape of the tumor from different angles so as to deliver full tumoricidal dose, while normal tissue is spared from excess radiation.

Figure 1. A linear accelerator used for external beam radiotherapy treatment.

In IMRT optimization, photon fluence from a beam is subdivided into "beamlets", which may be imagined to be divergent rectangular solids of fluence emanating from a radiation source in the LINAC's treatment head. One dimension of these beamlets, call it the "height", is defined by the projection of the MLC leaves onto a plane that is perpendicular to the central axis of the LINAC's beam and located at the rotational isocenter of the LINAC. These height projections are typically between 0.5 and 1.0 cm. In the "width" direction the resolution of the beamlet (projected on the same plane) is typically between 0.2 and 1.0 cm.

Optimization is over beamlets, whereas treatment delivery employs "beam segments" or "field segments", which are collections of beamlets that have been set to have the same intensity. The use of field segments is necessary for two reasons: (1) aggregations of many very small field dose calculations (i.e., on the order of a single beamlet) are extremely difficult, and (2) treatment time is proportional to the number of fields delivered. For reasons of economy and patient comfort treatment times are necessarily kept short.

Radiation dose, measured in Gray (Gy), is energy (Joules) deposited locally per unit mass (Kg). Fluence for external beam photon radiation is defined mathematically by the number of photon crossings per surface area. Dose tends to be proportional to fluence, but is influenced by photon and electron scatter in the patient's tissues as well as the energy and media involved. For any beam, selection of beamlet fluence weights results in a "fluence map" (intensity map) for that beam. Figure 2 shows the beam's-eye-view of a 7×7 beam with 44 of the 49 beamlets on. Different shades are used to reflect the different intensity of each beamlet.

A critical aspect of computer algorithms used in radiation therapy (or any other medical field) is that they must be fast enough that they do not impede the workflow of the clinic. Waiting for hours or days for an improved or highly accurate result is generally not possible. Most current optimization algorithms for IMRT treatment planning search the space of beamlet fluence weights only. This is because if other beam delivery parameters (e.g., field segments, couch angles, gantry angles, etc.) are incorporated into the optimization, the time required is prohibitive. The mixed integer programming ap-

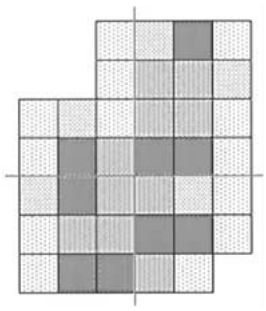

Figure 2. A beam's-eye-view of a 7×7 beam with 44 of the 49 beamlets on. Different shades are used to reflect the different intensity of each beamlet.

proach proposed herein allows the incorporation of such parameters simultaneously and yet returns solutions within minutes.

The planning process begins when the patient is diagnosed with a tumor mass and radiation is selected as part of the treatment regime. A 3D image, or volumetric studyset, of the affected region, which contains the tumor mass and the surrounding areas, is acquired via computed tomography (CT) scans. These CT data are used for treatment planning, and electron density information derived from it is used in the photon dose calculations for the beamlets. Additionally, magnetic resonance imaging (MRI) scans may be acquired, fused with the CT volumetric studyset, and used to identify the location and extents of some tumors – especially those in the brain. Based on these scans, the physician outlines the tumor and anatomic structures that need to be held to a low dose during treatment.

It is common practice to identify several regions of the tissue to be treated: The gross target volume (GTV) represents the volume which emcompasses the known macroscopic disease; that is, the disease that can be seen by the oncologist. The clinical target volume (CTV) expands the GTV to include regions of suspected microscopic disease. The planning target volume (PTV) includes additional margins for anatomical and patient setup uncertainties; that is, how the patient's organs and the patient will move from day to day. All volumetric data is discretized into voxels (volume elements) at a granularity that is conducive both to generating a realistic model and to ensuring that the resulting treatment planning integer programming instances are tractable (i.e., capable of being solved in a reasonable amount of computational time).

2.2. Dose calculation

The dose computation methods involve the principle of convolving the total energy release in the patient from the radiation source with Monte Carlo-generated enery deposition kernels and superposition of pencil beam (SPPB) using fundamental physics describing photon and electron interactions and transport. Our dose model accounts for the transport of primary and secondary radiation inside the patient, the variation of beam intensity across the patient surface, the effects of tissue inhomogeneities on the dose, and the irregular blocked or multi-leaf (MLC) shaped fields. The model consists of three components for computing the 3D dose distribution:

- Modeling the incident energy fluence as it exits the head of the linear accelerator.
- Projection of this incident fluence through the density representation of a patient to compute a Total Energy Released per unit MAss (TERMA) volume.
- A three-dimensional superposition of the TERMA with an energy deposition kernel using a ray-tracing technique to incorporate the effects of heterogeneities on lateral scatter.

The implementation was based on recent research in this area [1,12,13,22,28,29,32–34,36]. Figure 3 illustrates the process of the SPPB model for computing the dose to a

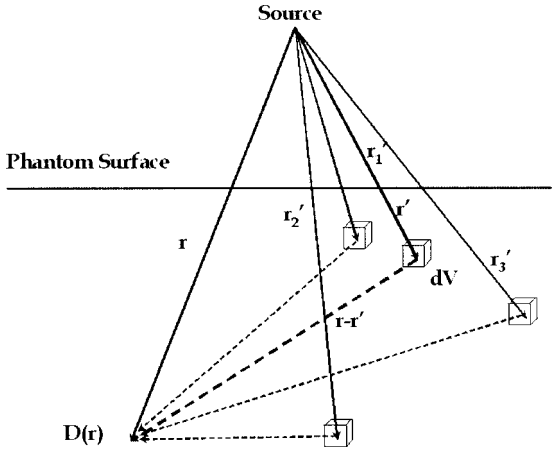

Figure 3. Calculation of dose from the shower of secondary particles resulting from primary intersection sites at r'.

point, $D(r)$. The dose at point $D(r)$ comprises contributions from the shower of secondary particles resulting from primary interaction sites at r'. The SPPB model provides accurate results within areas of electronic disequilibrium and tissue heterogeneities.

For each beamlet, the dose per intensity to a voxel is calculated using this dose engine. The total dose per intensity deposited to a voxel is equal to the sum of dose deposited from each beamlet. In this paper, for each patient, 16 non-coplanar candidate fields are generated. The size of the candidate fields and the associated number of beamlets is patient and tumor size dependent; varying from 10×10 cm^2 with 400 beamlets per field to 15×15 cm^2 with 900 beamlets per field. This results in a large set of candidate beamlets used for instantiating the treatment planning model.

2.3. Mixed integer programming treatment models

Let $\mathcal{B}$ denote the set of candidate beams, and let $\mathcal{N}_i$ denote the set of beamlets associated with beam $i \in \mathcal{B}$. Beamlets associated with a beam can only be used when the beam is chosen to be "on". If a beam is on, the beamlets with positive dose intensity will contribute a certain amount of radiation dosage to each voxel in the target volume and other anatomical structures. Once the set of potential beamlet intensities is specified, the total radiation dose received at each voxel can be modelled. Let $w_{ij} \geqslant 0$ denote the intensity of beamlet j from beam i. Then the total radiation dose at a voxel P is given by

$$\sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{N}_i} D_{P,ij} w_{ij}, \tag{1}$$

where $D_{P,ij}$ denotes the dose per intensity contribution to voxel P from beamlet j in beam i. Various dose constraints are involved in the design of treatment plans. Clinically

prescribed lower and upper bounds, say L_P and U_P, for dose at voxel P are incorporated with (1) to form the basic dosimetric constraints:

$$\sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{N}_i} D_{P,ij}w_{ij} \geqslant L_P \quad \text{and} \quad \sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{N}_i} D_{P,ij}w_{ij} \leqslant U_P. \tag{2}$$

Aside from constraining the dose received by each voxel within anatomic structures, we also constrain the number of beams used in the final beam profile. The motivation for this is that a simple plan (with a relatively small number of beams) is preferred over a more complex plan, since a complex plan takes more time to implement in the delivery room and offers more chances for errors. Let x_i be a binary variable denoting the use or non-use of beam i. The following constraints limit the total number of beams used in the final plan and ensure that beamlet intensities are zero for beams not chosen:

$$w_{ij} \leqslant M_i x_i \quad \text{and} \quad \sum_{i\in\mathcal{B}} x_i \leqslant B_{\max}. \tag{3}$$

Here, M_i is a positive constant which can be chosen as the largest possible intensity emitted from beam i, and $B_{\max}$ is the maximum number of beams desired in an optimal plan.

Dose-volume relationships within different anatomical structures are set up based on these constraints. Clinically, it is typically acceptable when 95% of the PTV receives the prescription dose, *PrDose*. The coverage constraints for PTV can thus be modeled as

$$\sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{N}_i} D_{P,ij}w_{ij} - r_P = PrDose, \quad P \in \text{PTV}, \tag{4}$$

$$r_P \leqslant D^{\text{OD}}_{\text{PTV}} v_P, \tag{5}$$

$$r_P \geqslant D^{\text{UD}}_{\text{PTV}}(v_P - 1), \tag{6}$$

$$\sum_{P\in\text{PTV}} v_P \geqslant \alpha|\text{PTV}|. \tag{7}$$

Here, v_P is a 0/1 variable which captures whether voxel P satisfies the prescription dose bounds or not; r_P is a real-valued variable that measures the discrepancy between prescription dose and actual dose; α corresponds to the minimum percentage of coverage required (e.g., $\alpha = 0.95$); $D^{\text{OD}}_{\text{PTV}}$ and $D^{\text{UD}}_{\text{PTV}}$ are the maximum overdose and maximum underdose levels tolerated for tumor cells; and |PTV| represents the total number of voxels used to represent the planning target volume. The values $D^{\text{OD}}_{\text{PTV}}$ and $D^{\text{UD}}_{\text{PTV}}$ must be chosen with care since inappropriately chosen values could cause the system of constraints to be infeasible.

It is desirable that dose received by organs/tissues other than the tumor volume should be minimal, as there is a direct correlation between the level of radiation exposure and normal tissue toxicity. Thus, for other anatomical structures involved in the planning process, along with the basic dose constraints given in (2), additional binary variables are

employed for modeling the dose–volume relationship. The dose–volume relationship is a standard metric that clinicians use when assessing a plan. It is a quantitative measure of the percentage volume of the anatomical structure receiving dose within specified intervals. To incorporate this concept into the model, let α_k, $\beta_k \in (0, 1]$ for k in some index set K, and let $y_P^{\alpha_k}$ and $z_P^{\alpha_k}$ be binary variables. Then the following set of constraints ensures that at least $100\beta_k\%$ of the voxels in an organ-at-risk, OAR, receive dose less than or equal to $\alpha_k\, PrDose$. In our models, the cardinality of the index set K is between 3 and 10.

$$\sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{N}_i} D_{P,ij}w_{ij} \leqslant [\alpha_k\, PrDose]y_P^{\alpha_k} + \mathcal{D}_{\max}z_P^{\alpha_k}, \quad P \in \text{OAR}, \tag{8}$$

$$\sum_{P\in\text{OAR}} y_P^{\alpha_k} \geqslant \beta_k|\text{OAR}|, \tag{9}$$

$$y_P^{\alpha_k} + z_P^{\alpha_k} = 1, \tag{10}$$

$$y_P^{\alpha_{k_1}} \leqslant y_P^{\alpha_{k_2}} \quad \text{for } \alpha_{k_1} \leqslant \alpha_{k_2}. \tag{11}$$

Here, $\mathcal{D}_{\max}$ is the maximum dose tolerance allowed for OAR, and α_k, β_k combinations are patient and tumor specific.

There are many objective functions that can be used to drive the optimization engine. For the computational work presented herein, we focus on the objective of minimizing a weighted sum of the excess dose to the PTV and the total dose to organs-at-risk. For comparisons of these and other models, readers are referred to [27].

3. Computational issues

The MIP instances include the basic dosimetric and volumetric constraints as described in (3)–(11) in addition to other clinical constraints. The resulting MIP instances have at least $\sum_{i\in\mathcal{B}}|N_i| + 1 + 3(|\text{PTV}| + 1) + \sum_{i\in\mathcal{O}}|K|(2|\text{OAR}_i| + 1) + (|K| - 1)|\text{OAR}_i|$ constraints; $\sum_{i\in\mathcal{B}}|N_i| + |\text{PTV}|$ continuous variables; and $|B| + |\text{PTV}| + \sum_{i\in\mathcal{O}} 2|K||\text{OAR}_i|$ binary variables, where $\mathcal{O}$ is the set of all organs-at-risk and normal structures. For real patient cases, there are tens of thousands of variables and constraints. For such cases, the instances have proven to be computationally very difficult for competitive commercial MIP solvers. Here, we describe a few specialized techniques that have been implemented to assist in the solution process.

3.1. Constraint and column generation

To maintain a tractable linear program relaxation, at a node of the branch-and-bound tree, instead of setting up the entire problem instance using all the voxel information, we generate a master problem which consists of roughly half of the original voxels. This subset is selected carefully in order to maintain a realistic description of the problem. As the solution process proceeds, additional voxels are introduced. This leads to the addition of constraints and the corresponding columns (variables). Constraints which

have remained inactive for a specified number of LP solves are removed from the master problem, thus providing a mechanism for controlling the size of the master instance. Interested readers can refer to [26] for clinical implications of the choice of voxels for modeling and the associated quality of the resulting treatment plans.

3.2. Specialized set-based branching scheme

For the constraint $\sum_{i\in\mathcal{B}} x_i \leqslant B_{\max}$ which bounds the number of beams (gantry angles and directions) selected in the final plan, instead of branching on each binary variable with fractional value, we branch on sets of binary variables. In particular, let x^{LP} be the fractional solution. The branching scheme partitions $\mathcal{B}$ into $\mathcal{B}_1 \cup \mathcal{B}_2$ such that $\sum_{i\in\mathcal{B}_1} x_i^{\mathrm{LP}}$ approximately equals $\sum_{i\in\mathcal{B}_2} x_i^{\mathrm{LP}}$. In addition, an attempt is made to choose each set B_i so that the included beams are roughly in the neighbourhood of each other. Two new nodes are then created via the constraints $\sum_{i\in\mathcal{B}_1} x_i \leqslant \lfloor B_{\max}/2 \rfloor$ and $\sum_{i\in\mathcal{B}_2} x_i \leqslant \lceil B_{\max}/2 \rceil$.

3.3. Geometric heuristics procedure

The heuristic procedure is an LP-based primal heuristic in which at each iteration, some binary variables are set to 1 and the corresponding linear program is resolved. The procedure terminates when the linear program returns an integer feasible solution or when it is infeasible. In the former case, reduced-cost fixing is performed at the root node, as well as locally on each of the branch-and-bound nodes.

The heuristic procedure focuses on the binary variables

$$q = \left(v_P, y_P^{\alpha_k}, z_P^{\alpha_k}\right)$$

from constraints (5)–(11). Given a fractional solution obtained from an LP relaxation at a node, let $\mathcal{U} = \{j\colon q_j^{\mathrm{LP}} = 1\}$, $\mathcal{F} = \{j\colon 0 < q_j^{\mathrm{LP}} < 1\}$, and $q^{\max} = \max\{q_j^{\mathrm{LP}}\colon\ j \in \mathcal{F}\}$. The heuristic works by first setting all binary variables in $\mathcal{U}$ to 1. Next, any variable in $\mathcal{F}$ for which the fractional value exceeds $q^{\max} - \varepsilon$ is set to 1, where ε is chosen dynamically with each fractional LP solution. Finally, it sets to 1 any variable corresponding to a voxel that is in a specified neighbourhood of a voxel for which the associated binary variable was already set to 1 in the previous two steps. The final step is based on the premise that if a voxel satisfies a certain dose bound, then all voxels in its neighbourhood should also satisfy the dose bound. The implementation requires a one-to-one mapping between the variables and the geometric locations of the associated voxels in a fixed 3D coordinate system.

3.4. Disjunctive cuts

In 1960 Gomory [18] first described a disjunctive argument to develop valid inequalities for mixed integer programs. In 1975 Balas [3] presented a general disjunctive approach which has been the basis of most recent computational research in this area [5–7,9,14, 25]. Disjunctive cuts, similar to Gomory cutting planes, have the appeal that they can

be applied to general integer programs without requiring any knowledge of the facial structure of the underlying polyhedron. Below, we describe our implementation.

Consider the polyhedron

$$P_{\text{IP}} = \text{conv}\{x \in \Re^n_+ \colon \hat{A}x \leqslant \hat{b},\ x_j \in \{0,1\},\ j = 1, \ldots, p\},$$

where $\hat{A}x \leqslant \hat{b}$ includes $Ax \leqslant b$ and the restrictions $x_j \leqslant 1$ for $j = 1, \ldots, p$; $\hat{A} \in \Re^{\bar{m} \times n}$. Let $x^t \in \Re^n_+$ be a feasible solution of $\hat{A}x \leqslant \hat{b}$ such that $0 < x^t_i < 1$ for some $i \in \{1, \ldots, p\}$ and consider the pair of polyhedra

$$P_{x_i,0} = \{x \in \Re^n_+ \colon \hat{A}x \leqslant \hat{b},\ x_i = 0\},$$
$$P_{x_i,1} = \{x \in \Re^n_+ \colon \hat{A}x \leqslant \hat{b},\ x_i = 1\}.$$

Clearly $P_{\text{IP}} \subseteq P_{x_i} \equiv \text{conv}(P_{x_i,0} \cup P_{x_i,1})$. Assume that both $P_{x_i,0}$ and $P_{x_i,1}$ are nonempty (otherwise, x_i, can be eliminated). The following fact, which is motivated by results in Balas [4], forms the basis of our cut-generation procedure.

Fact. The system

$$\begin{aligned}
&\hat{A}y - \hat{b}y_0 \leqslant 0,\\
&\hat{A}z - \hat{b}z_0 \leqslant 0,\\
&z_i - z_0 = 0,\\
&y_i = 0,\\
&z_0 + y_0 = 1,\\
&z + y = x^t,\\
&y, z, y_0, z_0 \geqslant 0
\end{aligned}$$

is infeasible if and only if $x^t \notin P_{x_i}$.

This together with Gale's Theorem of the Alternative [16,31] implies that $x^t \notin P_{x_i}$ if, and only if, the following linear system is feasible:

$$\begin{aligned}
&\alpha + \beta^T x^t < 0,\\
&u_1^T \hat{A} + u_4 e_i + \beta^T I \geqslant 0,\\
&u_2^T \hat{A} + u_3 e_i + \beta^T I \geqslant 0,\\
&-u_1^T b + \alpha \geqslant 0,\\
&-u_2^T b - u_3 + \alpha \geqslant 0,\\
&u_1, u_2 \geqslant 0,
\end{aligned}$$

where $u_1, u_2 \in \Re^{\bar{m}}$, $\beta \in \Re^n$, and $u_3, u_4, \alpha \in \Re$. From the latter system, form a linear program by (a) removing the first inequality and embedding it into the objective: $\min\{\alpha + \beta^T x^t\}$ and (b) enforcing an appropriate bounding condition on β. Such a linear

program will be referred to as a *disjunctive* LP. If the optimal objective value of a disjunctive LP is negative, then the inequality $\beta^T x \geqslant -\alpha$ is a valid inequality for P_{x_i} which cuts off the fractional solution x^t.

Empirical tests on the patient instances reveal that it is beneficial to generate cuts first based on the fractional variables $q = (v_P, y_P^{\alpha_k}, z_P^{\alpha_k})$. For each such 0/1 variable that satisfies $0.01 < q_i < 0.99$, we solve the corresponding disjunctive problem. In our implementation, $\|\beta\|_1 \leqslant 1$ (ℓ_1 norm) is used as the bounding condition for β. Computationally, this procedure is expensive as exactly one linear program of approximately twice the size of the original MIP instance must be solved. We perform this cut-generation procedure at the root node, as well as at tree levels that are a multiple of 10 within the branch-and-bound tree. To avoid excessive computational time, we select pseudo-randomly only 10% of the fractional variables for cut generation.

4. Numerical results

The numerical work reported in the remainder of the paper is based on a specialized branch-and-bound MIP solver which is built on top of a general-purpose mixed integer research code (MIPSOL) [24], using CPLEX V7.1 as the intermediate LP solver. The general-purpose code, which incorporates pre-processing, reduced-cost fixing, cut generation, and fast heuristics, has been shown to be effective in solving a wide variety of large-scale real-world MIP instances. Our algorithmic design thus far has been motivated by clinical cases. Special features as described in section 3 have been incorporated to assist in the solution process for solving these IMRT MIP instances.

We have tested our system on a collection of patient cases with tumor sites in various parts of the body. Here, we highlight five cases. For each case, several MIPs are solved, each having a different value for the maximum number beam angles ($B_{\max}$) allowed in the final plan, as imposed in constraint (3). In all cases, there are a total of 16 non-coplanar *candidate* beams each with 400 beamlets. A clinical comparison of plans associated with varying $B_{\max}$ is given in the next section. Here, we compare "MIP" statistics regarding timing, cuts generated, etc. For brevity, we only include results associated with $B_{\max} \in \{8, 12, 16\}$. More detail on implementation issues and comparisons of computational strategies will be reported in a companion computational paper.

Table 1 shows the sizes of the problem instances, including number of rows, number of columns and number of binary variables for each of the patient cases. Table 2 summarizes the solution statistics for one implementation strategy. *Initial LP Obj.* and *Optimal MIP Obj.* denote the objective values of the initial LP relaxation and the optimal objective value of the original MIP. *First feasible CPU secs* and *(First feasible) Obj* denote the time elapsed from the beginning of the solution process to when the algorithm returns an integer feasible solution, and the corresponding objective value. *Cut Number* and *Cut Time* denote the total number of cuts generated upon termination of the solution process, and the time used to generate the cuts. Finally, *Total elapsed CPU secs* denotes the number of CPU seconds needed to solve the instances to proven optimality.

Table 1
Problem statistics.

Pt	Rows	Columns	0/1 variables
1	27804	23820	14032
2	36946	32098	23110
3	54146	39280	29356
4	48092	42098	36134
5	54986	49182	41602

Table 2
Solution statistics.

Pt	Initial	First feasible		Optimal	Cut		Total elapsed
	LP obj.	CPU secs	Obj.	MIP obj.	Number	Time	CPU secs
				8 beams			
1	85934.6	402.6	135372.2	119202.6	230	4679.2	9178.4
2	168482.1	902.1	1311807.4	1120073.4	260	9201.5	14293.7
3	286829.3	1569.2	1911480.3	1594028.3	387	13290.7	25872.4
4	762383.4	1297.3	3541081.2	2801921.2	291	10275.3	23701.3
5	542102.3	1458.0	2920239.4	2190293.0	308	12015.1	29012.8
				12 beams			
1	84132.6	321.5	131892.4	112175.3	212	4012.3	9003.4
2	161465.1	802.7	1305027.6	1102156.2	278	10396.4	13639.2
3	280356.7	1321.4	1895023.7	1579396.1	310	12039.3	24803.0
4	729013.5	1045.2	3479372.1	2780123.2	328	14156.7	23201.3
5	518390.3	1301.6	2890273.4	2178016.5	289	10876.1	28112.4
				16 beams			
1	82376.0	210.7	130291.2	109320.8	221	4203.1	9012.1
2	154018.2	480.1	1290812.4	1078423.5	262	9012.3	13549.7
3	272623.1	927.5	1851560.3	1529039.4	345	13102.5	24193.6
4	702730.4	811.3	3420391.2	2750391.7	321	12907.6	22301.9
5	501390.7	1149.2	2830492.4	2174630.2	301	11245.0	27985.3

We observe that by way of construction, any solution from a 12-beam or 8-beam MIP is feasible for the 16-beam problem. This is reflected in the LP relaxation objective value and the optimal MIP objective value. As expected, the bulk of the CPU time is spent generating cutting planes to close the gap in the objective value. Overall, the total sum of radiation dose to normal tissue and organs-at-risk (the objective value) and the solution times appear to decrease with an increase in the number of beams used. On average, about 45% of the beamlets have positive weight in a feasible solution; moreover, the weights (intensities) vary significantly, indicating a high level of modulation in the resulting plan.

Table 3
Brain tumor case. Figures of merit for five plans distinguished by the number of allowed beams.

No. of beams	Coverage	Conformity	Homogeneity	Toxicity to C-shape
16	0.99	1.3	1.1	0.8
12	0.98	1.3	1.5	1.1
8	0.99	1.3	1.6	1.2
6	1.00	1.5	1.4	1.0
4	0.99	1.6	1.5	1.2

5. Clinical results

We include here results corresponding to a brain tumor case and a prostate tumor case. The plans presented correspond to the first *feasible* solution returned by our algorithm. Each plan is evaluated quantitatively based on isodose curves, dose–volume histogram, and four figures of merit: coverage index, conformity index, homogeneity index and the toxicity index. The former three indices are calculated according to the Radiation Therapy Oncology Group (RTOG) guidelines. Specifically, *coverage* is defined as the ratio of the tumor volume within the prescribed isodose surface to the total target volume. *Conformity* is defined as the ratio of the volume of the prescribed isodose surface to the target volume. *Homogeneity* is defined as the ratio of the maximum dose received by the tumor volume to the prescribed dose. Along with these RTOG indices, *toxicity* is used to measure radiation to organs-at-risk and normal tissue. The toxicity index is computed as the ratio of the maximum dose received by the specified proximal critical/normal tissue to the prescribed dose. A small toxicity score implies that the normal tissue does not receive an excessive amount of radiation. From the definitions of the four figures of merit, one can observe that these figures are not entirely independent. For example, while it is desirable to obtain a prescription isodose surface big enough to cover the target volume in order to ensure good coverage, it is also desirable to have this surface "small" in order to conform to the target volume. In addition, variations in conformity and coverage affect the amount of irradiation to nearby organs at risk, thus affecting toxicity levels of these organs.

The first case involves a metastatic melanoma brain tumor (white mass) with a spherical shaped $1.4 \times 1.9 \times 1.3$ cm^3 tumor located in the frontal lobe. The critical structure for this case is a C-shaped organ as outlined by the white curves in figure 4. Five plans derived from the MIP model by varying the number of allowed beams (4, 6, 8, 12 or 16 beams) were generated. Table 3 shows the figures of merit for each plan. Minimizing the excess dose to the tumor helps to obtain plans with good conformity and homogeneity in dose distribution. The dose restriction on critical structures helps to achieve low toxicity scores for the C-shaped critical structure.

Figures 4 and 5 contrast three orthogonal views (axial, sagittal, and coronal) of 16-beam and 6-beam IMRT plans obtained from the MIP model. In both figures, red represents the 100% prescription isodose curve, green represents the 70% isodose curve and blue represents the 50% isodose curve. Visually there is only a marginal difference

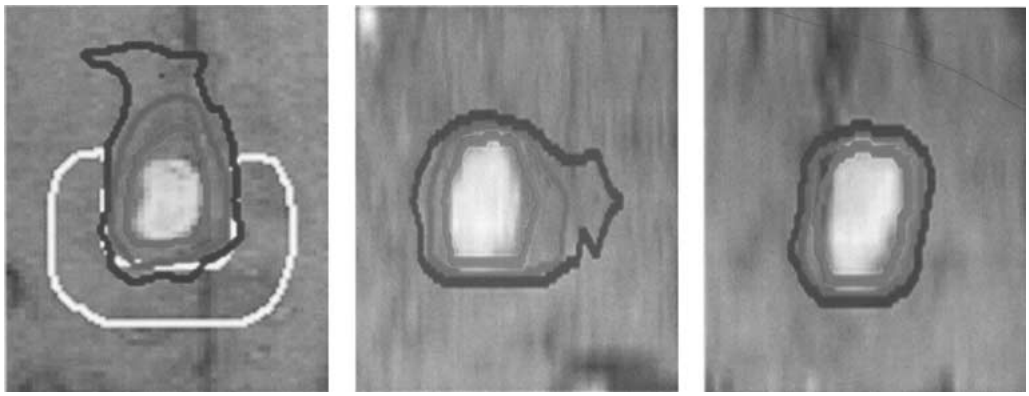

Figure 4. Brain tumor case. Isodose curves for IMRT plan with 16 beams. Axial, sagittal, and coronal views.

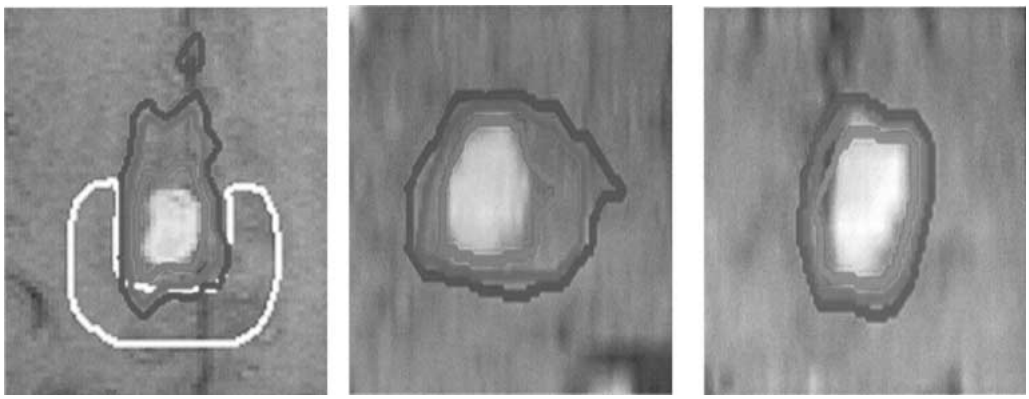

Figure 5. Brain tumor case. Isodose curves for IMRT plan with 6 beams. Axial, sagittal, and coronal views.

between these two plans. Quantitatively, we can observe from table 3 that when 16 beams are used, a more homogeneous dose and better conforming plan is delivered, while the C-shape structure is maintained at a low toxicity level.

The second case is an early stage prostate tumor. The critical structures here include the bladder and the rectum. Table 4 shows the figures of merit for each plan obtained. In viewing the dose–volume histogram in figure 6, observe the homogeneous dose delivered to the prostate. The rectum and bladder both receive slightly higher radiation when only 6 beams are used. This degradation in toxicity and conformity is observed consistently in table 4 when the number of beams used decreases. Figure 7 contrasts the isodose curves from the same two plans on one traversal view slice.

In general, clinical tests have shown that the computational engine can return good feasible solutions within 30 minutes, and the associated plans from these solutions provide good clinical figures of merit. Compared to certain commercial treatment planning systems, some consistent characteristics of plans resulting from the MIP models in-

Table 4
Prostate tumor case. Figures of merit for five plans distinguished by the number of allowed beams.

No. of beams	Coverage	Conformity	Homogeneity	Toxicity	
				to bladder	to rectum
16	0.99	1.12	1.100	1.042	1.042
12	0.98	1.24	1.108	1.126	1.101
8	0.99	1.35	1.111	1.140	1.125
6	1.00	1.46	1.116	1.145	1.126
4	0.99	1.46	1.148	1.175	1.155

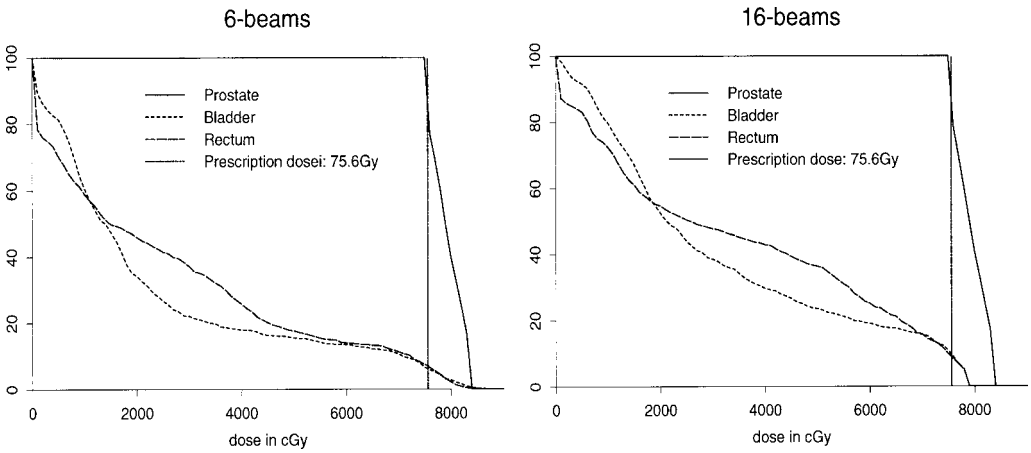

Figure 6. Prostate tumor case. Dose-volume histograms for two IMRT plans with 6 beams and 16 beams. The y-axis represents "percent volume greater than dose level". In both plans, homogeneous dose is delivered to the prostate. Note that there is virtually no difference in the dose received by the prostate. However, a lower maximum dose (toxicity) to both the rectum and bladder is observed with the 16-beams plan.

clude (a) superior dose homogeneity over the tumor volume, (b) reduction of radiation to organs-at-risk and nearby normal tissues, and (c) improvement in conformity while maintaining the desired tumor coverage. The results provide evidence that the MIP approach is viable in producing superior treatment plans which can potentially lead to significant improvement in local tumor control and reduction in normal tissue complication. In addition, our tests demonstrate that real-time planning is achievable.

6. Conclusion

A novel integer programming approach for intensity modulated treatment planning optimization has been presented. The MIP model proposed allows simultaneous optimization over the space of beamlet fluence weights and beam and couch angles. Based on our experiments with clinical data, this approach can return good plans which are clinically acceptable and practical. The plans consistently provide homogeneous and conformal

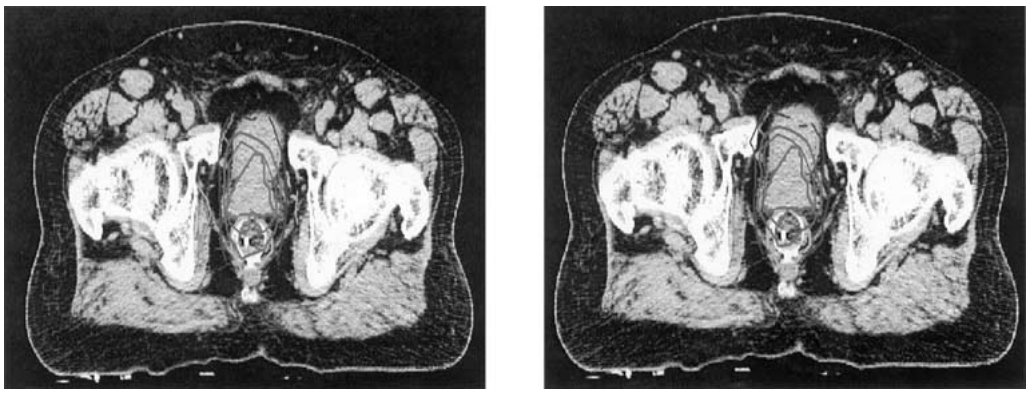

Figure 7. Prostate tumor case. Isodose curves for IMRT plan with 6 beams and 16 beams, respectively. The prostate is outlined in **pink**. The two critical structures are bladder (top **yellow** contour) and rectum (low yellow contour). **Red** represents the 100% prescription isodose curve, **green** represents the 70% isodose curve and **blue** represents the 50% isodose curve. Note the more conformal dose in the 16-beam plan (right).

dose to the tumor, while maintaining low irradiation to critical structures. Although the mixed integer programming instances are difficult to solve to optimality, the specialized techniques implemented enable solving them to proven-optimality. On average, the first feasible solution is returned within 30 minutes, and is of high quality clinically. Compared to currently available systems, most of which perform optimization over only on a subset of beam parameters, this MIP approach allows consideration of a more comprehensive set of parameters; and with the reasonable solution time, it is viable for incorporation within a real-time treatment planning system.

Acknowledgment

This research was partially supported by the National Science Foundation. Special thanks are extended to Dr. Mark Wiesmeyer of Computerized Medical System Inc., for his time in discussion of general IMRT treatment in commercial systems and his insightful comments on an earlier version of this manuscript. Finally, we thank the anonymous reviewers for their comments for improving the manuscript.

References

[1] A. Ahnesj, Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media, Medical Physics 16 (1989) 577–592.
[2] G. Bahr, J. Kereiakes, H. Horwitz, R. Finney, J. Galvin and K. Goode, The method of linear programming applied to radiation treatment planning, Radiology 91 (1968) 686–693.
[3] E. Balas, Disjunctive programming: cutting planes from logical conditions, in: *Nonlinear Programming*, Vol. 2, eds. O.L. Mangasarian et al. (Academic Press, New York, 1975) pp. 279–312.
[4] E. Balas, Disjunctive programming, Annals of Discrete Mathematics 5 (1979) 3–51.

[5] E. Balas, S. Ceria and G. Cornuéjols, A lift-and-project cutting plane algorithm for mixed 0/1 programs, Mathematical Programming 58 (1993) 295–324.
[6] E. Balas, S. Ceria and G. Cornuéjols, Mixed 0–1 programming by lift-and-project in a branch-and-cut framework, Management Science 42 (1996) 1229–1246.
[7] E. Balas, A modified lift-and-project procedure, Mathematical Programming 79 (1997) 19–32.
[8] J. Barbiere, M.F. Chan, J. Mechalakos, D. Cann, K. Schupak and C. Burman, A parameter optimization algorithm for intensity-modulated radiotherapy prostate treatment planning, J. Appl. Clin. Medical Physics 3(3) (2002) 227–234.
[9] R.E. Bixby, W. Cook, A. Cox and E.K. Lee, Computational experience with parallel mixed integer programming in a distributed environment, Annals of Operations Research 90 (1995) 19–43.
[10] T. Bortfeld, K. Jokivarsi, M. Goitein, J. Kung and S.B. Jiang, Effects of intra-fraction motion on IMRT dose delivery: statistical analysis and simulation, Physics in Medicine and Biology 47(13) (2002) 2203–2220.
[11] T. Bortfeld, U. Oelfke and S. Nill, What is the optimum leaf width of a multileaf collimator? Medical Physics 27(11) (2000) 2494–2502.
[12] **?**. Bourland and **?**. Chaney, A finite-size pencil beam model for photon dose calculations in three dimensions, Medical Physics 19 (1992) 1401–1412.
[13] A.L. Boyer and E.G. Mok, A photon dose distribution model employing convolution calculations, Medical Physics 12 (1985) 169–177.
[14] S. Ceria and G. Pataki, Solving integer and disjunctive programs by lift and project, in: *Integer Programming and Combinatorial Optimization, Proceedings of the 6th International IPCO Conference*, eds. R.E. Bixby et al., Lecture Notes in Computer Science, Vol. 1412 (Springer, Berlin, 1998) pp. 271–283.
[15] S.M. Crooks and L. Xing, Linear algebraic methods applied to intensity modulated radiation therapy, Physics in Medicine and Biology 46(10) (2001) 2587–2606.
[16] D. Gale, *The Theory of Linear Economic Models* (McGraw-Hill, New York, 1960).
[17] R.E. Gomory, Solving linear programs in integers, in: *Combinatorial Analysis*, eds. R.E. Bellman and M. Hall, Jr. (American Mathematical Society, Providence, RI, 1960), pp. 211–216.
[18] R.E. Gomory, An algorithm for the mixed integer problem, RM-2597, The Rand Corporation, 1960.
[19] T. Fox, A dose calculation engine for an intensity-modulated treatment planning system, Working Paper (2002).
[20] M. Hartmann, L. Bogner, M. Fippel, J. Scherer and S. Scherer, IMCO(++) – a Monte Carlo based IMRT system, Medical Physics 12(2) (2002) 97–108.
[21] M. Hilbig, R. Hanne, P. Kneschaurek, F. Zimmermann and A. Schweikard, Design of an inverse planning system for radiotherapy using linear optimization, Medical Physics 12(2) (2002) 89–96.
[22] P.W. Hoban, D.C. Murray and W.H. Round, Photon beam convolution using polyenergetic energy deposition kernels, Physics in Medicine and Biology 39 (1994) 669–685.
[23] M. Langer, R. Brown, M. Urie, J. Leong, M. Stracher and J. Shapiro, Large scale optimization of beam weights under dose-volume restrictions, International Journal of Radiation Oncology and Biological Physics 18 (1990) 887–893.
[24] E.K. Lee, Computational experience of a general purpose mixed 0/1 integer programming solver (MIPSOL), Technical Report, Georgia Institute of Technology (1997).
[25] E.K. Lee, Generating cutting planes for mixed integer programming problems in a parallel computing environment, Technical Report, Georgia Institute of Technology, to appear in INFORMS Journal on Computing (2000).
[26] E.K. Lee, T. Fox and I. Crocker, Effects of beam configuration and tumor representation on dosimetry and plan quality, Medical Physics, in review (2002).
[27] E.K. Lee, T. Fox and I. Crocker, Sensitivity analysis of clinical objectives to plan qualities in intensity modulated radiation therapy treatment planning optimization, Medical Physics, in review (2002).
[28] T.R. Mackie, J.W. Scrimger and J.J. Battista, A convolution method of calculating dose for 15 MV X rays, Medical Physics 12 (1985) 188–196.

[29] T.R. Mackie, A.F. Bielajew, D.W.O. Rogers and J.J. Battista, Generation of photon energy deposition kernels using the EGS Monte Carlo code, Physics in Medicine and Biology 33 (1988) 1–20.
[30] P.E. Metcalfe, P.W. Hoban, D.C. Murray and W.H. Round, Beam hardening of 10 MV radiotherapy X-rays: analysis using a convolution/superposition method, Physics in Medicine and Biology 35 (1990) 1533–1549.
[31] O.L. Mangasarian, *Nonlinear Programming* (McGraw-Hill, New York, 1959).
[32] R. Mohan, C. Chui and L. Lidofsky, Energy and angular distributions of photons from medical linear accelerators, Medical Physics 12 (1985) 592–597.
[33] R. Mohan, C. Chui and L. Lidofsky, Differential pencil beam dose computation model for photons, Medical Physics 13 (1986) 64–73.
[34] N. Papanikolaou, T.R. Mackie, C. Meger-Wells, M. Gehring and P. Reckwerdt, Investigation of the convolution method for polyenergetic spectra, Medical Physics 20 (1993) 1327–1336.
[35] A.B. Pugachev, A.L. Boyer and L. Xing, Beam orientation optimization in intensity-modulated radiation treatment planning, Medical Physics 27(6) (2000) 1238–1245.
[36] O.Z. Ostapiak, J. Van Dyk and Y. Zhu, A model for 3D photon dose calculations based on finite size pencil beams generated using FFT convolutions, AAPM Annual Meeting Program, Medical Physics 22 (1995) 976.
[37] R.J. Schulz and A.R. Kagan, On the role of intensity-modulated radiation therapy in radiation oncology, Medical Physics 29(7) (2002) 1473–1482.