Radiotherapy optimAl Design: An Academic Radiotherapy Treatment Design System

R. Acosta[a*], W. Brick[b**], A. Hanna[c***], A. Holder[d***], D. Lara[e**],
G. McQuilen[b*], D. Nevin[f*], P. Uhlig[g***], B. Salter[h]

June 14, 2008

Abstract

Optimally designing radiotherapy and radiosurgery treatments to increase the likelihood of a successful recovery from cancer is an important application of operations research. Researchers have been hindered by the lack of academic software that supports head-to-head comparisons of different techniques, and this article addresses the inherent difficulties of designing and implementing an academic treatment planning system. In particular, this article details the algorithms and the software design of Radiotherapy optimAl Design ($\mathcal{RAD}$).

Keywords: Optimization, Radiotherapy, Radiosurgery, Medical Physics.

[a] Stanford University, Department of Computational Mathematics, Palo Alto, CA

[b] Trinity University Mathematics, San Antonio, TX

[c] St Mary's University, Department of Computer Science, San Antonio, TX

[d] Rose-Hulman Institute of Technology, Department of Mathematics, Terre Haute, IN

[e] Altarum, San Antonio, TX

[f] Texas A & M University, Department of Industrial Engineering, College Station, TX

[g] St. Mary's University, Department of Mathematics, San Antonio, TX

[h] The University of Utah, Department of Radiation Oncology, Salt Lake City, UT

[*] Research supported by NCI-SURF grant #6 R25 CA080936-06.

[**] Research supported by The Cancer Therapy & Research Center, San Antonio, TX.

[***] Research supported by the Department of Radiation Oncology, Huntsman Cancer Center.

1 Introduction

With the USA spending about 18% of its gross national product on health care, the need to efficiently manage and deliver health services has never been greater. In fact, some prominent researchers have claimed that if we are not judicious with our resources, then our health care system will burden society with undue costs and vast disparities in availability [10, 24]. Developing mathematical models that allow us to study and optimize medical treatments is crucial to the overall goal of efficiently managing the health care industry. Indeed, we have already witnessed medical advances by optimizing existing medical procedures, leading to better patient care, increased probability of success, and better time management.

Much of the previous work focuses on using operations research to improve administrative decisions, but several medical procedures are now being considered. The breadth and importance of these is staggering, and the academic community is poised to not only aid managerial decisions but also improve medical procedures. One of the most prominent examples of the latter is the use of optimization to design radiotherapy treatments, which is the focus of this article.

Cancer remedies largely fall into three categories: pharmaceutical - such as chemotherapy, surgical - whose intent is to physically remove cancerous tissues, and radiobiological - which uses radiation to destroy cancerous growths. Radiotherapy is based on the fact that cancerous cells are altered by radiation in a way that prevents them from replicating with damaged DNA. When a cell is irradiated with a beam of radiation, a secondary reaction forms a free radical that damages cellular material. If the damage is not too severe, a healthy cell can likely overcome the damage and replicate normally, but if the cell is cancerous, it is unlikely that it will be able to regenerate. The differing abilities of cancerous and non-cancerous cells to repair themselves is called a therapeutic advantage, and the goal of radiotherapy is to deliver enough radiation so that cancerous cells expire but not so much as to permanently damage nearby tissues.

Radiotherapy treatments are delivered by focusing high energy beams of ionizing radiation on a patient. The goal of the design process is to select the pathways along which the radiation will pass through the anatomy and to decide the amount of dose that will be delivered along each pathway, called a *fluence value*. Bahr [6] first suggested that we optimize treatments in 1968, and since then medical physicists have developed a plethora of models to investigate the design process. Currently, commercially available planning systems optimize only the fluence value and do not additionally optimize the pathways. All of these commercially available systems rely on some form of optimization algorithm and these algorithms can range from gradient descent to simulated annealing. To date, the optimization approaches implemented clinically, typically by medical physicists working in the field of radiation oncology, have been reasonably effective but have failed to exploit the significant advances of robust operations research theory. As operations researchers have become aware of such problems, increasingly sophisticated optimization expertise has been brought to bear on the problem, leading to a growing potential for more elegant solutions.

The knowledge barrier between medical physicists, who understand the challenges and nuances of treatment design, and operations researchers, who know little about the clinical environment, is problematic. Clinical capabilities vary significantly, making what is optimal dependent on a specific clinic (treatments also depend on an individual patient). So, the separation of knowledge stems not only from the fact that the operations researchers generally know little about medical physics, but also from the fact that they typically know even less about the capabilities of a specific clinic. This lack of understanding is being overcome by several collaborations between the two communities, allowing academic advances to translate into improved patient care.

One of the most significant research hindrances is the lack of head-to-head comparisons, with the vast majority of numerical calculations being undertaken by individual research groups on patients from their associated clinic. This work describes the academic treatment system $\mathcal{R}$adiotherapy optim$\mathcal{A}$l $\mathcal{D}$esign ($\mathcal{RAD}$), which is designed to use

- standard optimization software to model and solve problems,
- a database to store cases in a well defined manner, and
- a web based interface for visualization.

The use of standard modeling software makes it simple to alter and/or implement new models, a fact that supports head-to-head comparisons of the various models suggested in the literature. Storing problems in a database is an initial step toward creating a test bank that can be used by researchers to make head-to-head comparisons, and the web interface facilitates use. These features agree with standard OR practice in which algorithms and models are compared on the same problems and on the same machine.

The corresponding author once believed that a rudimentary version of $\mathcal{RAD}$ was possible within a year's effort. This was an extremely naive thought. $\mathcal{RAD}$'s implementation began 1999, with the initial code being written in Matlab. The current version is written in C++ and PHP and links to AMPL, CPLEX and a MySQL database. At every turn there were numerical difficulties, software engineering obstacles, and verification problems with the radiation transport model. The current version required the concentrated effort of eight mathematics/computer science students, three Ph.Ds in mathematics/computer science, and one Ph.D in Medical Physics spread over 3 years. The details of our efforts are contained herein.

1.1 The Nature of Radiotherapy

Radiotherapy is delivered by immobilizing a patient on a horizontal table, around which a linear accelerator or gantry, capable of producing a beam of radiation, rotates. The table may be moved vertically and horizontally, allowing the central point of gantry rotation, called the *isocenter*, to be placed anywhere in the anatomy, and may be rotated in a horizontal plane. The gantry is capable of rotating 360 degrees around the patient, although some positions are not allowed due to collision potential, see Figure 1.

Treatment design is typically approached in 3 phases,

1) **Beam Selection** Select the pathways through which the radiation will pass through the anatomy.

2) **Fluence Optimization** Decide how much radiation (fluence) to deliver along each of the selected beams to best treat the patient.

3) **Delivery Optimization** Decide how to deliver the treatment computed in the first two phases as efficiently as possible.

The first two phases of treatment design are repeated in the clinic as follows. A designer uses sophisticated image software to visually select beams (also called pathways or angles) that appear promising. The fluence to deliver along these beams is decided by an optimization routine, and the resulting treatment is judged. If the treatment is unacceptable, the collection of beams is updated and new fluences are calculated. This trial-and-error approach can take as much as several hours per patient. Once an acceptable treatment is created, an automated routine decides how to sequence the delivery efficiently. There is an inherit disagreement between the objectives of fluence and delivery optimization since a fluence objective improves as beams are added but a delivery objective degrades. Extending the delivery time is problematic since this increases the probability of patient movement and the likelihood of an inaccurate delivery.

The initial interest in optimizing radiotherapy treatments was focused on fluence optimization, and numerous models and solution procedures have been proposed,

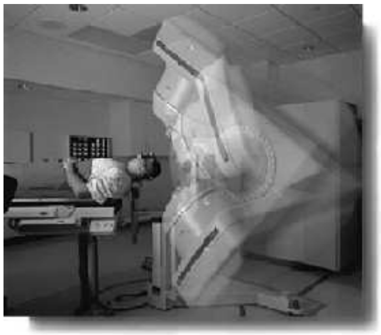

Figure 1: A standard treatment delivery system.

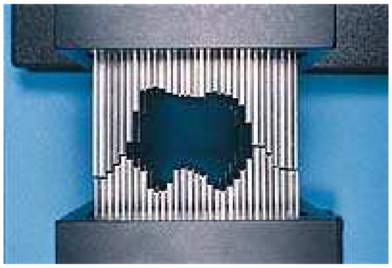

Figure 2: A multileaf collimator.

see [7, 17, 18, 25, 26] for reviews. The variety is wide and includes linear, quadratic, mixed integer linear, mixed integer quadratic, and (non) convex global problems. Clinically relevant fluence problems are large enough to make combining the first two phases, which is trivial to express mathematically, difficult to solve, and much of the current research is directed at numerical procedures to support a combined model. One of $\mathcal{RAD}$'s strengths is that it is designed so that different models and solution procedures can easily be compared on the same cases, allowing head-to-head comparisons that were previously unavailable.

There are many treatment paradigms, with one of the most common being Intensity Modulated Radiotherapy (IMRT). This treatment modality is defined by the use of a multileaf collimator housed in the head of the gantry. The leaves of the collimator shape the beam, see Figure 2, and by adjusting the beam's shape continuously during treatment, we can modulate the delivered dose. The idea is to adjust the leaves so that different parts of the anatomy receive variable amounts of radiation. By combining several beams from multiple orientations, we hope to deliver a uniform tumoricidal dose while sparing the surrounding healthy organs and tissues.

While the treatment advantages of a collimator are apparent, the collimator significantly adds to the complexity of treatment design since it provides the ability to control of small subdivisions of the beam. This is accomplished by dividing the open-field into sub-beams, whose size is determined by the collimator. For example, the collimator in Figure 2 has 32 opposing leaves that vertically divide the open-field. Although the leaves move continuously, we horizontally divide the open-field to approximate the continuous motion and design a treatment that has a unique value for each rectangular sub-beam. The exposure pattern formed by the sub-beams is called the fluence pattern, and an active area of research is to decide how to best adjust the collimator to achieve the fluence pattern as efficiently as possible [1, 4, 9].

A radiotherapy treatment is designed once at the beginning of treatment, but the total dose is delivered in multiple daily treatments called fractions. Fractionization allows normal cells the time to repair themselves while accumulating damage to tumorous tissues. The prescribed dose is typically delivered in 20 to 30 uniform treatments. Patient re-alignment is crucial, and in fact, the beam of radiation can often be focused with greater precision than a patient can consistently be re-aligned. Radiosurgery treatments differ from radiotherapy treatments in that the total intended dose is delivered all at once in one large fraction. The intent of a radiosurgery is to destroy, or ablate, tissue. Patient alignment is even more important for radiosurgeries because the large amount of radiation being delivered makes it imperative that the treatment be delivered as planned.

2 Calculating Delivered Dose

The radiation transport model that calculates how radiation energy per unit mass is deposited into the anatomy, which is called dose, is crucial to our ability to estimate the anatomical effect of external beam radiation. Obviously, if the model that describes the deposition of dose into the patient does not accurately represent the delivered (or anatomical) dose, then an optimization model that aids the design process is not meaningful.

Numerous radiation transport models have been suggested, with the "gold standard" being a stochastic model that depends on a Monte Carlo simulation. This technique estimates the random interactions between the patient's cells and the beam's photons, and although they are highly accurate, such models generally require prohibitive amounts of computational time (although this is becoming less of a concern). We instead adapt the deterministic model from [23] that approximates each sub-beam's dose contribution. The primary dose relies on the beam's energy and on the ratio between the radius of the sub-beam and the open-field. The way a sub-beam scatters as it travels through the anatomy depends on the radius of the sub-beam. Small radius beams have a large surface area compared to their volume, and hence, they lose a greater percentage of their photons than do larger radius sub-beams. When many contiguous sub-beams are used in conjunction, much of this scatter is gained by surrounding sub-beams, called scatter dose buildup.

The radiation transport model estimates the rate at which radiation is deposited, and we let $D_{(p,a,i)}$ be the rate at which dose point p gains radiation from sub-beam i in angle a. A few comments on the term 'dose point' are warranted. Much of the literature divides the anatomy into 3D rectangles called voxels and then considers the amount of radiation delivered to an entire voxel. This approach is well suited to other radiobiological models, but the one that we use is based on geometric distances. To calculate these distances, we represent each voxel by its center and call this a dose point. The units of $D_{(p,a,i)}$ are Grays per fluence, where one Gray is equal to 1 joule per kilogram. The triple (p, a, i) defines the depth d of dose point p along sub-beam (a, i), see Figure 3. Beams *attenuate* as they pass through the anatomy, meaning that lose energy as they pass through tissue. The maximum accumulation is not at the anatomy's surface but rather at a depth of M due to the previously mentioned dose buildup, after which the attenuation is modeled as exponential decay, $e^{-\mu(d-M)}$, where μ is an energy parameter. For a 6MV beam, the maximum dose rate is typically observed to occur at a depth of 1.5 cm. The dose rate at the surface is approximately 60% of the maximum rate at depth M, and we linearly interpolate between this value at depth 0 and the maximum rate at depth M. While this interpolation is not exact, it is reasonable and the majority of critical structures are at depths greater than M. The dose model we use is

$$D_{(p,a,i)} = \begin{cases} \left(P_0\, e^{-\mu(d-M)}(1-e^{-\gamma r}) + \frac{rd\alpha_d}{r+M}\right) \times ISF \times O, & d \geq M \\ \\ \left(\frac{0.4d}{M} + 0.6\right) \times \left(P_0\,(1-e^{-\gamma r}) + \frac{rM\alpha_M}{r+M}\right) \times ISF \times O, & 0 \leq d < M. \end{cases}$$

The primary dose contribution for depths at least M is $P_0\, e^{-\mu(d-M)}$, where P_0 is a machine-dependent constant. The factor $(1 - e^{-\gamma r})$ represents the percentage of the open-field radiation present in a sub-beam, where γ is experimentally defined and r is the radius of the sub-beam. Notice that as r decreases the sub-beam's intensity falls exponentially, so extremely small sub-beams are not expected to effectively treat deep tissue malignancies. The term $(rd\alpha_d)/(r + M)$ models the scatter contributions from the surrounding sub-beams, where

$$\alpha_d = -0.0306 \ln(d) + 0.1299.$$

Again, the contribution from scatter decreases with r (although linearly instead of exponentially).

The final two factors are the inverse square factor, ISF, and the off-axis factor, O. The inverse square factor is the square of the ratio of the distance from the gantry to the isocenter and the distance from the gantry to the dose point. Allowing

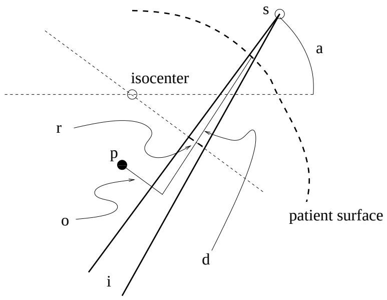

Figure 3: The geometry of a radiobiological model.

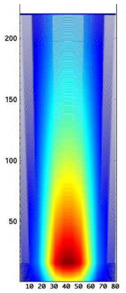

Figure 4: The dose contour of a single sub-beam.

$l(s, \text{isocenter})$ and $l(s,p)$ to be the distances from the source s to the isocenter and from s to the dose point p, we have

$$ISF = \left(\frac{l(s, \text{isocenter})}{l(s,p)}\right)^2.$$

The off-axis factor adjusts the dose contribution so that dose points near a sub-beam accumulate radiation faster than those that are farther away. This factor depends on the off axis distance o in Figure 3 and is machine and beam energy dependent, making it necessary to use a table tailored to a specific machine. Linear interpolation is performed to determine off axis contributions for distances not listed in the table.

For distances greater than M, Nizin [23] reports that the maximum error is 5% for clinically relevant beams when compared to Monte-Carlo models. For extremely narrow beams, which are not clinically relevant, there is a maximum error of 8%. For our purposes, this level of accuracy is sufficient.

The dose rates discussed in the previous section are arranged to form a dose matrix, denoted by A, whose rows are indexed by p and whose columns are index by (a,i). Allowing $x_{(a,i)}$ to be the fluence of sub-beam i in angle a, the cumulative dose at point p is

$$\sum_{(a,i)} D_{(p,a,i)} x_{(a,i)} = (Ax)_p,$$

where the indices of the vector x correspond to the columns of A. So, the linear map $x \mapsto Ax$ transforms a fluence pattern into anatomical dose (Gy).

Although Figure 3 depicts angle a being divided into ‘flat’ sub-beams, the collimator segments a beam into a 2D grid. The column widths of the grid are decided by the width of the leaves, and the row widths depend on the type of collimator. Some collimators are binary, and each leaf is either open or closed. Other collimators allow each leaf to move continuously across the column, and in this situation the rows of the grid are used to discretize this continuous motion. The subscript i indexes through this grid at each angle, and hence, i is actually a 2D index. Similarly, the index for angles need not be restricted to a single great circle around the patient, and the index a represents an angle on a sphere around the patient.

3 Coordinate Systems

A complete foray into the authors' tribulations with the different coordinate systems is beyond the scope of this article, but a few notes are important. There are three coordinate systems that need to be aligned, 1) the coordinates for the patient images, 2) the coordinates for the dose points, and 3) the location of the gantry. The patient images are three dimensional of the form (μ, ν, ζ), where each ζ represents a typical cross sectional image. The images are not necessarily evenly spaced, with images often being closer through the target. The dose points are also three dimensional of the form $p = (u, v, z)$. As discussed below, placement of these points is restricted to an underlying, evenly space grid, and hence, the z coordinate do not necessarily agree with ζ. However, each dose point needs to be linked to a tissue that is defined by an image, and we associate (u, v, z) with (u, v, ζ_z), where ζ_z is the closest ζ to z. The gantry coordinates describe the machine and not the anatomy. To link the gantry's position and rotation to the patient, we need to know the location of the isocenter within the patient and the couch angle. Gantry coordinates are calculated in polar form and translated to rectangular coordinates that are synced with the anatomy's position on the couch.

Our solution to aligning the coordinate systems is to build a three dimensional rectangle whose coordinates are fixed and whose center is always the location of the isocenter. We load the patient images into the rectangle so that they position the isocenter accordingly, and then build a three dimensional grid in the rectangle that defines where dose points are allowed to be placed. The couch angle defines a great circle around the fixed rectangle that allows us to position and rotate the gantry, which in turn allows us to track the sub-beams as they gantry moves.

4 Optimization Models

The linear operator $x \mapsto Ax$ facilitates an optimization model. For convenience, we sub-divide the rows of A into those that correspond to the target, critical structures, and normal tissue, and we let A_T, A_C, and A_N be the corresponding sub-matrices so that A_Tx, A_Cx and A_Nx are the anatomical doses for the target, the critical structures, and the normal tissue. Physicians form a prescription by placing bounds and goals on the anatomical dose. The precise definition of a prescription depends on the definition of an optimal treatment. For example, a simple least squares model is

$$\min\{\omega_1\|A_Tx - TG\|_2 + \omega_2\|A_Cx\|_2 + \omega_3\|A_Nx\|_2 : x \geq 0\}. \tag{1}$$

The prescription for this model is the vector of goal doses for the targeted region, TG, which is commonly a scalar multiple of the vector of ones. The ω scalars weight the terms of the objective to express clinical preference. A linear model with a more complicated prescription is

$$\begin{aligned}\min\{\omega_1 \cdot \alpha + \omega_2 \cdot \beta + \omega_3 \cdot \gamma : TLB - \alpha e \leq A_Tx \leq TUB,\ A_Cx \leq CUB + \beta e,\\ A_Nx \leq NUB + \gamma e,\ 0 \leq A_Tx \leq TLB,\ \beta \geq -\| - CUB\|_\infty,\ \gamma \geq 0\},\end{aligned} \tag{2}$$

where e is the vector of ones (length is decided by the context of its use). The prescription for this model is TLB - vector of lower bounds for the target, TUB - vector of upper bounds for the target, CUB - vector of upper bounds for the critical structures, NUB - vector of upper bounds for the normal tissues, and the weighting values ω_i, $i = 1, 2, 3$. Both (1) and (2) are fluence problems since they define an optimal treatment for a known set of angles, sub-beams and dose points. In general, a prescription is the information provided by a physician that is used to define an instance of a fluence problem. Table 1 shows the prescription information gathered by $\mathcal{RAD}$. The $ABSMIN$ and $ABSMAX$ information is used in some models to define 'hard' constraints instead of the goals suggested by other bounds. For example, the model below differentiates between the goal lower bound TLB and the target's absolute minimum bound $ABSMIN_T$.

$$\min\{\|z\|_2 : ABSMIN_T \leq A_Tx \leq TUB{+}z,\ A_Cx \leq CUB{+}z,\ A_Nx \leq NUB{+}z,\ x \geq 0\}.$$

The hard constraint $ABSMIN_T \leq A_Tx$ guarantees the target will receive at least the values in $ABSMIN_T$, whereas TUB, CUB and NUB are goals parametrized by z.

	Notation	Description
Goals	TG	A goal dose for a target.
	TLB	A goal lower bound for a target.
	TUB	A goal upper bound for a target.
	CUB	A goal upper bound for a critical structure.
	NUB	A goal upper bound for the normal tissues.
Dose Bounds	$ABSMAX$	An absolute maximum dose allowed on any structure.
	$ABSMIN$	An absolute minimum dose allowed on the target.
Error Bounds	$PERABV$	Percent of volume allowed to violate an upper bound.
	$PERBLW$	Percent of volume allowed to violate a lower bound.

Table 1: Prescription information gathered by $\mathcal{RAD}$. Each of these vectors is indexed to accommodate the required number of structures.

Error bound parameters are commonly used in what are called dose volume constraints, which limit the volume of tissue allowed to violate a goal. For example, a mixed integer extension of (1) that limits the amount of under irradiated target is

$$\begin{aligned}&\min\{\omega_1\|A_Tx - TG\|_2 + \omega_2\|A_Cx\|_2 + \omega_3\|A_Nx\|_2 :\\ &\quad (1-y_p)\textstyle\sum_{(a,i)} A_{(p,a,i)}x_{(a,i)} \le TLB,\ \sum_{p\in T} y_p \le \text{PERBLW}\cdot|T|,\ y\in\{0,1\}^{|T|},\ x\ge 0\}.\end{aligned}$$

These are notoriously difficult problems with a substantial literature, see [13, 19] as examples.

Once an optimization model is decided, an optimal treatment can be calculated using a host of different solvers. If the model is linear, solver options include the primal simplex method, the dual simplex method, Lemke's algorithm, and several interior point methods. Unless the optimization problem has a unique solution, it is likely that different solution algorithms will terminate with different fluence patterns (although the objective values will be the same). This phenomena has been observed in [14], where CPLEX's dual simplex was shown to routinely design treatments with a few angles having unusually high fluences. If the model is nonlinear but smooth, typical options are gradient descent, Newton's method, and variants of quasi-Newton's methods.

In the spirit of $\mathcal{RAD}$'s academic intent, one of our goals is to allow easy and seamless flexibility in how optimal treatments are defined and calculated. This is possible by using pre-established software that is designed to model and solve an optimization problem. In particular, we separate data acquisition, modeling, and solving. This differs from the philosophy behind most of the in-house systems developed by individual clinics, where modeling and solving are intertwined in a single computational effort. The mingling of the two complicates the creative process because changing either the model or the solution procedure often requires large portions of code to be rewritten, thus hindering exploration. We instead use standard software to model and solve a problem. For instance, $\mathcal{RAD}$ uses AMPL to model problems, which makes adjusting existing models and entering new ones simple. AMPL links to a suite of 35 solvers such as (integer) linear and (integer) quadratic models (as well as many others). $\mathcal{RAD}$ currently has access to CPLEX, MINOS and PCx. This approach takes advantage of the numerous research careers that have gone into developing state-of-the-art software to model and solve optimization problems, and hence brings a substantial amount of the field of Operations Research to the design of radiotherapy treatments.

There are limitations to this design approach, especially with global (non-convex) problems that are often successfully solved with simulated annealing or other heuristics. The lack of access to quality heuristics for global optimization problems is a detriment because one of the industry's primary solution methods is simulated annealing. Simulated annealing has the favorable quality that it can successfully accommodate any model, but the unfavorable quality that there is no guarantee of optimality. However, the medical physics community has learned to trust this technique because it has consistently provided quality treatments. Moreover, some of the suggested optimization models are non-convex global models. A future goal is to link $\mathcal{RAD}$ to global solvers like LGO, simulated annealing, and genetic algorithms. Once this is complete, a large scale study of which model and solution methodology provides the best clinical benefit is possible. Wide scale testing of different models and solution procedures has not been undertaken, but $\mathcal{RAD}$ has the potential to support this work. Such work is important because clinical desires vary from clinic to clinic and from physician to physician. This leads to the situation where the sense of optimality —i.e. the optimization model, can be different from one physician to another for the same patient. It is possible, however, that the basic treatment goals pertinent to all clinics and physicians are best addressed by a single model and solver combination. If this is true, then such a model and solver combination would provide a consensus on what an optimal treatment is for a specific type of cancer and a subsequent standard of care for clinics with similar technology.

5 Problem Management

The size of a typical design problem is substantial, making them difficult to solve. Indeed, the results in [12] show that storing the dose matrix can require over 600 Gigabytes of memory. For this reason, it is necessary to use reduction schemes to control the size of a problem, and this section discusses the methods introduced in $\mathcal{RAD}$, several of which are designed to assist the combination of beam selection and fluence optimization.

The current practice of asking the treatment planner to select beams has the favorable quality that the underlying fluence problem only considers the selected beams, which reduces the number of columns in the dose matrix A. $\mathcal{RAD}$ is capable of addressing a fluence problem with a large number of candidate beams by judiciously selecting sub-beams and dose points. The first reduction is to simply remove the sub-beams that do not strike the target. One naive way to remove these sub-beams is to calculate an entire dose matrix and then remove the columns whose aggregate rate to the target is below a specified threshold. $\mathcal{RAD}$'s approach is different, and before calculating the rates associate with a sub-beam, we search for the minimum off-axis factor over the dose points on the surface of the target. If the minimum value is too great, we classify the sub-beam as non-target-striking and are spared the calculation of this column. This technique requires two calculations that are not needed by the naive approach, that of locating the target's surface and evaluating the minimum off-axis factor. Both of these calculations only consider the target, whereas the naive approach calculates rate information for each point in the anatomy. Our numerical comparisons, even for large targets, show that $\mathcal{RAD}$'s approach is significantly faster.

A novel reduction introduced in $\mathcal{RAD}$ is to accurately define the anatomical region that will receive significant dose. For example, consider a lesion in the upper part of the cranium. The volume defined by the entire set of patient images is called the *patient volume*, and for this example it would likely encompass the head and neck. However, it is unlikely that we will need to calculate dose in the neck region. We have developed a technique that defines a (typically much) smaller volume within the patient where meaningful dose is likely to accumulate.

Assume that a designer considers a unique great circle around the patient by selecting a single isocenter and couch angle, represented by the pair (c, j). Further assume that beams will be selected from the collection of beams placed every 5^o on this great circle. Using only the sub-beams that strike the target, we trace them through the anatomy to define a swath. A mathematical description highlights the ease with which this region can be calculated. Let B_0 be the plane defined by the gantry as it rotates around the isocenter with a fixed couch angle. B_0 is defined by the unit normal vector N and the isocenter c_0: $B_0 = \{c_0 + N\nu : N\nu = 0\}$. Two additional planes B_1 and B_2 are defined using the same normal vector N, so that they are parallel to B_0, but with the respectively different points,

$$\begin{aligned} c_1 &= \left(\max\left\{\text{dist}(E(i), B_0) : i \in S\right\} + r\right) N \\ c_2 &= \left(\min\left\{\text{dist}(E(i), B_0) : i \in S\right\} - r\right) N. \end{aligned}$$

In this calculation, S is the set of all sub-beams that are target striking and E is the map from S into $\mathbb{R}^3$ so that $E(i)$ is the point where sub-beam i exits the anatomy. The distance between this point and the B_0 plane is

$$\text{dist}(E(i), B_0) = \frac{N^T E(i) - N^T c_0}{\|N\|}.$$

Note that this is the signed minimum distance between a point and a plane. Points in the direction of N have a positive distance, and points in the direction of $-N$ have a negative distance. Allowing D to be the collection of all possible dose points in the patient volume defined by the entire set of images, we define the swath for isocenter c and couch angle j to be

$$W_{(c,j)} = \{d \in D : \text{dist}(d, B_1) \leq 0\} \cap \{d \in D : \text{dist}(d, B_2) \geq 0\}.$$

This development shows that constructing the swath is computationally simple because we only need to iterate over the elements of S, which are already defined by the first reduction, calculate $\text{dist}(E(i), B_0)$, and keep the maximum and minimum elements. We additionally add any delineated critical structures that lay outside this region. The combined region is called the *treatment volume* for (c, j). If there are further isocenters and couch angles, the entire treatment volume is the union over all (c, j). Dose points are only placed within this volume for planning purposes, reducing the number of rows of the dose matrix.

The arrangement of dose points over the treatment volume is critical for two reasons: 1) the discrete representation of the anatomical dose needs to accurately estimate the true continuous anatomical dose, and 2) the size of the problem grows as dose points are added. Clinical relevance is achieved once the dose points are within 2mm of each other. Some technologies are capable of taking advantage of 0.1mm differences, and hence, require much finer grids. Our experiments show that using the treatment volume instead of the patient volume reduces the storage requirement to 10s of Gigabytes instead of 100s of Gigabytes with a grid size of 2 mm, assuming a single isocenter and couch angle. Although this is a significant reduction, solving a linear or quadratic problem with a dose matrix in the 10 Gigabyte range is impossible with CPLEX on a 32 bit PC, there are simply not enough memory addresses.

Our final reduction is a technique that iteratively builds a dose matrix for the treatment volume. The idea is to eliminate the numerous dose points in the normal tissue that are largely needed to protect against a *hot spot*, which is clinically defined as a cubic centimeter of tissue receiving an unusually high dose, say 110% of the average target dose. While hot spots are not desired in any part of the anatomy, including the target, the general consensus is that hot spots should never be located outside the target. Our final reduction scheme attempts to ensure that we place normal tissue dose points so that we can monitor areas likely to have hot spots while eliminating the dose points in the normal tissue that do not influence the problem. $\mathcal{RAD}$ uses the following iterative procedure:

1. On the first solve we only include normal tissue dose points that are adjacent to the target, forming a *collar* around the target. This concept was used in the earliest work in the field [5].

2. We segment the patient volume into 1cm^3 sectors.

3. We trace the sub-beams that have sufficiently high fluence values, and each sector is scored by counting the number of high fluence sub-beams that pass through it.

4. Normal tissue dose points are placed within the treatment volume for the sectors that have high scores.

5. The process repeats with the added dose points until every sector receives a sufficiently small score.

This iterative procedure solves several small problems instead of one large one. On clinical examples, the initial dose matrices are normally under 1 Gigabyte, a size that is appropriate for the other software packages. We point out that $\mathcal{RAD}$ does not calculate the anatomical dose to each sector but rather only counts where high exposure sub-beams intersect. Just because a few high exposure sub-beams pass through a sector does not mean that this sector is a hot spot, but it does mean that it is possible. Sectors with low counts can not be hot spots because it is impossible to accumulate enough dose without several high exposure sub-beams. We find that 1 to 5 iterations completely removes hot spots outside the target. At the end of the process, the dose matrix has typically grown negligibly and remains around 1 Gigabyte in size.

The iterative procedure above reduces the number of rows of the dose matrix so dramatically and successfully that we can increase the number of beams. Although a clinic will only use a fraction of the added beams, solving for optimal fluences with many beams provides information about which beams should and should not be selected. A complete development of beam selection is not within the scope of this work, and we direct readers to [14] for a complete development. Beam selectors are classified as *uninformed*, *weakly informed* or *informed*. An uninformed selector is one that only uses geometry, and the current clinical paradigm of selecting beams by what looks geometrically correct is an example. A weakly informed selector is guided by the dose matrix and the prescription, and an informed selector further takes advantage of an optimal fluence pattern calculated with a large set of possible beams. The premise behind an informed selector is that it begins with a large set of possible beams that are allowed to 'compete' for fluence through an optimization process. The expectation is that a beam with a high fluence is more important than a beam with a low fluence. The numerical results in [14] demonstrate that informed selectors usually select quality beams in a timely fashion.

$\mathcal{RAD}$ is designed to study beam selection and fluence optimization and is well positioned to address the current research pursuit of simultaneously optimizing both fluence and beams [2, 3, 20]. Rather than asking a user to identify beams, users are instead asked to select collections of couch angles and isocenters, which is clinically easier. $\mathcal{RAD}$ then places angles on each of the great circles, calculates the treatment volume, and uses the iterative procedure above to control hot spots. Beams are typically spaced at 5^o increments on each of the great circles. The resulting optimal treatment is not clinically viable but is available for an informed beam selector. Of course, uninformed and weakly informed beam selectors are possible as well, but $\mathcal{RAD}$ provides the additional option of using an informed selector. Once beams are selected, $\mathcal{RAD}$ places dose points on a fine mesh (spacing no greater than 2 mm) throughout the treatment volume. Since the number of beams is small, this leads to a dose matrix that remains appropriate with our other software packages. A final optimal treatment is calculated with this matrix. Table 2 contains the expected size reductions for a 20cm^3 region.

We conclude this section with a brief discussion about sparse matrix formats and other reductions that did not work. A straight forward approach of reducing the storage requirements of the dose matrix is to store only those values above a predefined threshold. This method requires the calculation of every possible rate coefficient over the patient volume. Our approach of defining the treatment volume preempts the majority of these calculations and is faster. That said, about 90% of the rate coefficients over the treatment volume are insignificant since each sub-beam

Size	Number of Rows	Number of Columns	Size of A
Patient Volume	10^6	1.3×10^5	1.3×10^{11}
Sequential Solves	3.0×10^4	8.4×10^2	2.5×10^7
Final Solve	3.8×10^5	4.0×10	1.5×10^7

Table 2: Approximate dose matrix sizes for a 20cm^3 region with 3 couch angles around a single isocenter in the middle of the patient volume. A 2mm 3D grid spacing is assumed. Each swath is 1cm in width (6 dose points wide), is parallel to one of the axes, and passes through the center of the patient volume. The example assumes that $10,000$ dose points in the treatment volume are not normal and that $20,000$ additional dose points outside the treatment volume are needed to describe the critical structures. Each beam is assumed to have a 25×25 grid of sub-beams, of which 4 are assumed to strike the target. The final treatment has 10 angles.

delivers the majority of its energy to a narrow path through the treatment volume. So, a sparse matrix format over the treatment volume should further reduce our storage requirements. However, our reduction schemes allow us to forgo a sparse matrix format and store a dose matrix as a simple 2D array with double precision. This simplicity has helped us debug and validate the code.

Before arriving at the reductions above, we attempted a different method of placing dose points. The idea was to use increasingly sparse grids for the target, critical structures, and normal tissues. This is not a new idea, with different densities being considered in [21, 22]. There are two problems with this approach. First, the voxels of different grids have different volumes, and our code to handle the volumes at the interface of different grids was inefficient. Second, the sparsity of the normal tissue grid had to exceed 1 cm (often 2+ cm) to accommodate the use of many angles. This is well beyond clinically acceptable values. Moreover, the sparsity did not prevent hot spots from appearing in the normal tissue. We are aware that Nomos's commercial software uses an octree technique that allows varying densities, so it is possible to use this idea successfully, although our attempt failed.

6 Solution Analysis

A treatment undergoes several evaluations once it is designed. In fact, the number of ways a treatment is judged is at the heart of this research, for if the clinicians could tell us exactly what they wanted to optimize, we could focus on optimizing that quantity.

However, no evaluative tool comprehensively measures treatment quality, making the problem inherently multiple objective [15, 16]. The issue is further complicated by the fact that treatment desires are tailored to specific patients at a specific clinic. That said, there are two general evaluative tools.

A Dose Volume Histogram (DVH) is a graph that for each structure plots dose against percent volume, which allows a user to quickly evaluate how the entirety of each structure fairs. A treatment and its corresponding DVH from the commercial system Nomos are found in Figures 5 and 6. The upper most curve of the DVH corresponds to the target, which is near the brain stem. This curve starts to decrease at about 52 Gy, which indicates that 100% of the target receives at least this dose. The next highest curve represents the spinal cord, a structure whose desired upper bound is 45 Gy. This curve shows that about 18% of the spinal cord is to receive a higher dose. The remaining curves correspond with the eye sockets and remaining normal tissue.

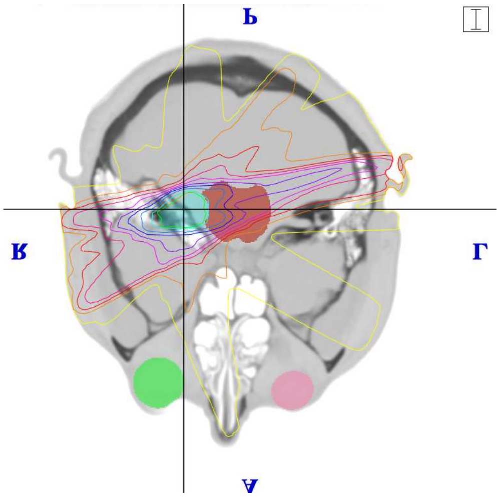

Figure 5: The isodose contours for a clinically designed treatment.

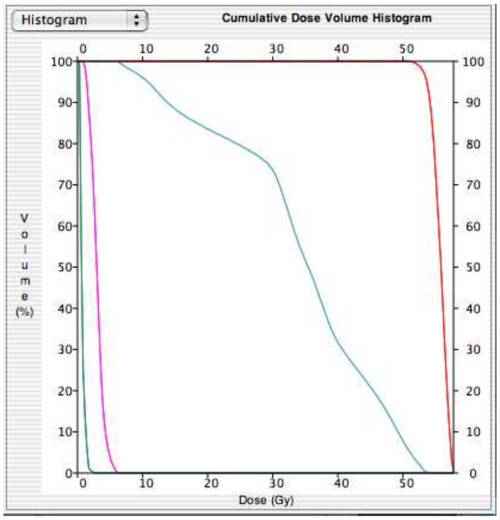

Figure 6: The DVH for the treatment in Figure 5.

Notice that a DVH curve depends on the volumetric estimate of the corresponding structure, an observation that leads to a subtle issue. Different clinics are likely to create different volumes of the normal tissue by scanning different patient volumes. This means the curve for normal tissue will vary, and in particular, the information provided by this curve diminishes as more normal tissue is included. For example, if we are treating the lesion in Figure 5, we could artificially make it appear as though less than 1% of the normal tissue receives a significant dose by including the tissue in the patient's feet. The authors of this paper are unaware of any standard, and for consistency, all of $\mathcal{RAD}$'s DVHs are based on the treatment volume, which is a definable and reproducible quantity that removes the dependence on the clinically defined patient volume.

A DVH visually displays the amount of a structure that violates the prescription but does not capture the spacial position of the violation. If 10% of a structure's volume violates a bound but is distributed throughout the structure, then there is likely no radiobiological concern. However, if the volume is localized, then it might be a hot spot and the treatment is questionable. To gain spatial awareness of where dose is and is not accumulated, each of the patient images is contoured with a sequence of *isodose curves*. Examples are found in Figure 5. Each of these curves contains the region receiving dose above a specified percentage of the target dose. So, a 90% isodose curve contains the tissue that receives at least $0.9 \times TG$. Isodose curves clearly indicate the spatial location of high dose regions, but they require the user to investigate each image and form a mental picture of the 3D dose. Since there are often hundreds of scans, this is a tedious process.

Rendering isodose curves proved more complicated than we had first assumed, and $\mathcal{RAD}$ incorporates a new approach. We build a cubic approximation of the continuous dose with a B-spline. Let $\delta_{(k,t)}$, $\delta_{(k+1,t)}$, $\delta_{(k,t+1)}$ and $\delta_{(k+1,t+1)}$ be the dose at four neighboring dose points on one of the patient images (recall that dose on a image is associated with the dose at the closest dose point - since we are dealing with a single image, we remove the ζ coordinate). The cubic approximation over this region is

$$S_{(k,t)}(\mu,\nu) = (1/36)UMQ_{(k,t)}M^TV^T, 0 \le \mu,\nu, \le 1,$$

where $U = [\mu^3, \mu^2, \mu, 1], V = [\nu^3, \nu^2, \nu, 1]$,

$$M = \begin{bmatrix} -1 & 3 & -3 & 1 \\ 3 & 6 & 3 & 0 \\ -3 & 0 & 3 & 0 \\ 1 & 4 & 1 & 0 \end{bmatrix} \text{ and } Q_{(k,t)} = \begin{bmatrix} \delta_{(k-1,t-1)} & \delta_{(k-1,t)} & \delta_{(k-1,t+1)} & \delta_{(k-1,t+2)} \\ \delta_{(k,t-1)} & \delta_{(k,t)} & \delta_{(k,t+1)} & \delta_{(k,t+2)} \\ \delta_{(k+1,t-1)} & \delta_{(k+1,t)} & \delta_{(k+1,t+1)} & \delta_{(k+1,t+2)} \\ \delta_{(k+2,t-1)} & \delta_{(k+2,t)} & \delta_{(k+2,t+1)} & \delta_{(k+2,t+2)} \end{bmatrix}.$$

By design, these regional approximations combine to form a smooth approximation over the entire patient image, which is

$$S(\mu,\nu) = S_{(\lfloor\mu\rfloor,\lfloor\nu\rfloor)}(\mu - \lfloor\mu\rfloor, \nu - \lfloor\nu\rfloor),$$

where $(\mu, \nu) \in [1, m] \times [1, n]$. If the indexing exceeds the image, then exterior dose values are interpreted as the dose of their nearest neighbor. For example, $\delta_{(0,0)} = \delta_{(1,0)} = \delta_{(0,1)} = \delta_{(1,1)}$, of which $\delta_{(1,1)}$ is the only real dose value.

This is a traditional (uniform) B-spline, and nothing is new about continuously approximating discrete information with this technique. However, the cubic estimation allows us to draw an isodose curve by finding a level curve of $S(\mu, \nu)$ –i.e. the solutions to $S(\mu, \nu) - \alpha TG = 0$, where $0 \leq \alpha \leq 1$. What is new is that we use a shake-and-bake algorithm to identify the isodose curve [8]. The idea is to start within the region of high-dose, randomly select a direction, and then find the high dose boundary along the forward and backward rays. Formally, fix α between 0 and 1 so that we are looking for the α isocontour. Let (μ_0, ν_0) be a position on the patient image such that the dose $\delta_{(\mu_0,\nu_0)}$ is greater than αTG. Uniformly select ρ_0 in $[0, \pi)$, and along the line segment $(\mu_0, \nu_0) + \theta(\cos(\rho), \sin(\rho))$, find the smallest positive θ and largest negative θ such that either $S((\mu_0, \nu_0) + \theta(\cos(\rho_0), \sin(\rho_0))) - \alpha TG = 0$ or θ is at a bound that prevents the coordinate from leaving the patient image. This calculation renders a line segment defined by $\theta_0^{\max}$ and $\theta_0^{\min}$, and the next iteration begins with the midpoint of this segment,

$$(\mu_1, \nu_1) = (\mu_0, \nu_0) + (1/2)(\theta_0^{\max} + \theta_0^{\min})(\cos(\rho_0), \sin(\rho_0)).$$

This technique has the favorable mathematical property that if the region within the isocontour is convex, then the random sequence $(\mu_k, \nu_k) + \theta_k^{\max(\text{or } \min)}(\cos(\rho_k), \sin(\rho_k))$ converges to the uniform distribution on the isocontour [8]. Moreover, it is suspected, although not proved, that the uniformity is achieved for any connected region [11]. An example is shown in Figure 7.

We let (μ_0, ν_0) be the location of the dose point with the maximum dose. We mention that this may or may not be the largest value of $S(\mu, \nu)$, and another option would be to solve $\nabla S(\mu, \nu) = 0$ to find its maximum value. We use Newton's Method, with a full step, to solve $S((\mu_0, \nu_0) + \theta(\cos(\rho_0), \sin(\rho_0))) - \alpha TG = 0$, which has favorable quadratic convergence and is simple to implement since the partials of $S(\mu, \nu)$ are

$$\frac{\partial}{\partial \mu} S_{(k,t)}(u, v) = (1/36) U \hat{I} M Q_{(k,t)} M^T V^T$$

and

$$\frac{\partial}{\partial \nu} S_{(k,t)}(u, v) = (1/36) U M Q_{(k,t)} M^T \hat{I}^T V^T,$$

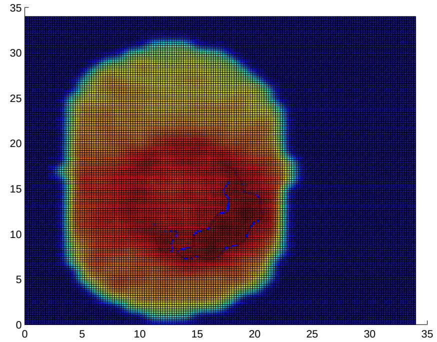

Figure 7: A 90% isocontour rendered with the shake-and-bake algorithm.

where

$$\hat{I} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Many alternative line searches are available, such as a binary search, and such techniques may provide stability if the gradient of S is near zero. Lastly, we mention that this method renders an unordered collection of points on the isocontour, and without an order, it is not clear how to connect them. This is not a concern if enough points are rendered, for after all, every displayed curve is a collection of pixels.

7 Software Design & Structure Identification

The previous sections' discussions about the algorithms that support $\mathcal{RAD}$ do not address the software engineering aspects, and the authors would be remiss if they did not discuss how the different parts of $\mathcal{RAD}$ interlink. Some of the topics in this section are general software issues and others are specific to the design of radiotherapy treatments.

The basic idea behind our approach is to exploit a language's strengths. As Table 2 indicates, the number of calculations needed to form a dose matrix is significant, and this part of the project is written in C++. Dose matrices may be written to disc for debugging purposes or they may be kept in memory and passed directly to other applications. Reading and writing a 1 Gigabyte file from and to disc is time consuming, and the latter approach saves time, especially when everything stays in

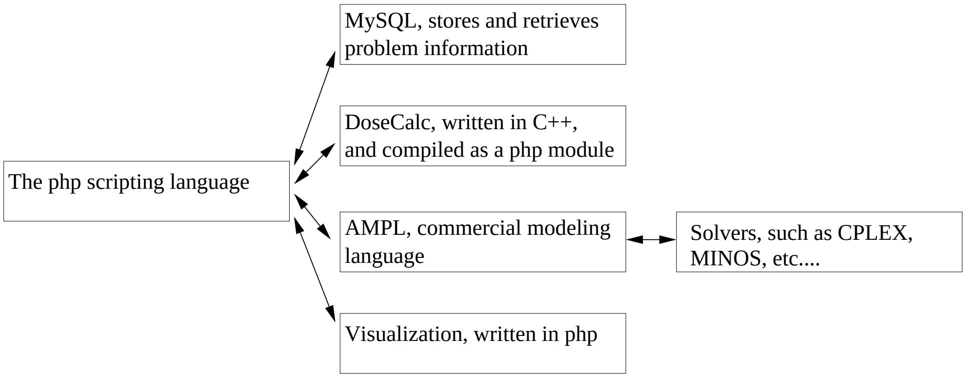
MySQL, stores and retrieves problem information
DoseCalc, written in C++, and compiled as a php module
The php scripting language
AMPL, commercial modeling language
Solvers, such as CPLEX, MINOS, etc....
Visualization, written in php

Figure 8: The SWIG compiler links PHP to C++, which in turn allows us to pass information between DoseCalc (the parallel implementation of our radio-biological model) and other software.

RAM. We use the SWIG compiler to link our C++ code to the scripting language PHP, which allows the dose matrix to be stored in a native PHP structure. Through the use of a bi-directional pipe, we can pass the dose matrix to other applications like AMPL and receive information when a problem is solved. Moreover, PHP gives us control of the linux command line, naturally interfaces with MySQL, and easily interfaces with the web to support the user interface. For these reasons, PHP became the ‘glue’ that held the system together, and the power of this ability should not be under estimated. Figure 8 depicts the general design.

Another of $\mathcal{RAD}$’s unique features is that it stores problems in a MySQL database. Beyond being an information repository for $\mathcal{RAD}$, its intent is to become a library of problems for head-to-head comparisons. The medical literature on treatment design is vast, but each paper highlights a technique on a few examples from a specific clinic, examples that can not be used by others to compare results. This is strange from a computational perspective, and $\mathcal{RAD}$’s database will support the numerical work necessary to fairly evaluate different algorithms and models.

A natural question about data storage is, What is a problem? In other words, what data is needed to define a problem instance. We adhere to the following structure,

$$\begin{aligned} \text{Problem} &= (\text{Case, Prescription, Settings}) \\ \text{Solution} &= (\text{Problem, Model, Solution Technique}). \end{aligned}$$

A Case is defined by the geometry of the patient. We parse RTOG files (a standard protocol for radiotherapy treatments) to gain a description of the structures that were delineated on the patient images. Each structure is defined by a series of polygons on individual patient images, with each polygon being defined by an ordered list of points. We construct a continuous boundary for each structure by linearly joining consecutive points on each image. We mention that other splining techniques were tested in an attempt to make the structural boundaries smooth, but none were more accurate when overlaid on the CT scan. Some structures, such as the kidney shown in Figures 9 and 10, may be described by several polygons on the same slice, creating geometries that complicate the process of automatically defining the regions associated with each tissue. The regions defined by these curves may or may not be contained within each other. If the regions are disjoint, we assume they enclose the same tissue. However, if one of the regions is within the other, such as in Figure 9, we assume the region defined by the inner curve is not part of the defined tissue. To test whether or not the regions are disjoint, we calculate a winding number. Let (μ_k, ν_k) and $(\hat{\mu}_k, \hat{\nu}_k)$ be two lists of points on the same image for the same structure. To see whether or not the polygon defined by $(\hat{\mu}_k, \hat{\nu}_k)$ is within the polygon defined by (μ_k, ν_k), we select a single $(\hat{\mu}_K, \hat{\nu}_K)$ and consider the vertical line through this point. For every directed line segment from (μ_k, ν_k) to (μ_{k+1}, ν_{k+1}) that passes the vertical line above (below) $(\hat{\mu}_K, \hat{\nu}_K)$ from right to left, we accrue 1 (-1). The signs reverses if we pass from left to right. In the event that (μ_{k+1}, ν_{k+1}) lies on the vertical line, we instead consider the directed line segment from (μ_k, ν_k) to (μ_{k+2}, ν_{k+2}) (or an even larger index if the second point is also on the vertical line) for the calculation. Under the assumption that regions are either disjoint or nested, which is an assumption we make, this calculation returns 0 if and only if the polygon defined by $(\hat{\mu}_k, \hat{\nu}_k)$ is within the polygon defined by (μ_k, ν_k).

Tissue information is captured with a tga image that is generated for each patient image by flooding each tissue with a unique color. For example, the three segments in Figure 10 would be linearly interpolated and the pixels within the outer polygon but outside the inner polygons would be flooded with a color unique to kidney tissue. This is possible with a PHP class that generates tga images, which are not stored but rather generated as needed from the list of points in the database (this saves storage requirements). Representative tga images for each tissue are displayed via a web interface that additionally asks the user for the prescription information for each tissue. Each dose point is associated with the closest pixel on a patient image, where ties are decided with a least index rule. Thus, the tga images are the link between the user defined prescription and the associated bounds of the optimization problem.

Another concern about tissue identification is that regions representing different tissues may intersect. Our simple solution follows that of several commercial systems, and we ask the user to rank tissues. In the case of an intersection, the dose points

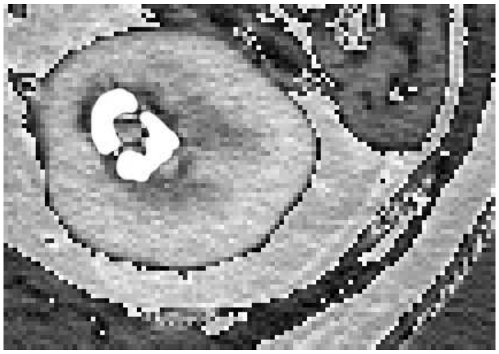

Figure 9: An enlarged view of a kidney. The white areas are not kidney and were delineated on the patient images.

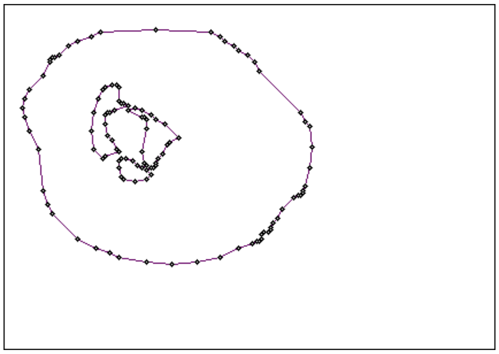

Figure 10: The kidney was defined for this patient image by three segments: an outer segment delineating the larger volume and two inner segments that contain non-kidney tissue. The points listed in the RTOG file are indicated by small circles.

within the intersection are labeled as the tissue with the highest priority.

We have already defined the information available for the prescription in Table 1. The *Settings* information details the dose point grid, the location of the possible angles, and the type & sub-division of the beam. Whereas the information that comprises a Problem details what is needed to design a treatment, a Solution additionally includes the type of optimization model and the technique used to solve it -i.e. it includes how we are defining and finding optimality. Hence, a Solution is everything needed to define the anatomical dose of an optimal treatment, and with this information it is possible to render a treatment to be evaluated.

8 Conclusion

Orchestrating the creation of a radiotherapy design system is a significant task that lives at the application edge of operations research and computer science. This paper has discussed many of the fundamental concerns and has introduced several new tactics that allow the underlying optimization problem to be approached with standard software, even in the case of numerous possible angles. It additionally introduces a new method to draw isocontours. The system is based on an efficient and well studied radiation transport model.

Many researchers have faced the challenge of designing their own design software, which is why there are several in-house, research systems. Our goal was to detail the algorithmic and software perspectives of $\mathcal{RAD}$ so that others can incorporate our experience as they either begin or continue their research. In the future, the authors will initiate the process of a rigorous, detailed and wide-spread investigation into which model and solution method consistently produces quality treatments. Moreover, $\mathcal{RAD}$ will allow others to compare their (new) techniques to ours with the same dose model and patient information. Lastly, $\mathcal{RAD}$ is designed to accommodate the amalgamation of beam selection and fluence optimization, a topic that is currently receiving interest.

References

[1] R.K. Ahuja and H.W. Hamacher. Linear time network flow algorithm to minimize beam-on-time for unconstrained multileaf collimator problems in cancer radiation therapy. Technical report, Department of Industrial and Systems Engineering, University of Florida, 2004. Revised version under review in Networks.

[2] D. Aleman, H. Romeijn, and J. Dempsey. A response surface approach to beam orientation optimization in intensity modulated radiation therapy treatment planning. In *IIE Conference Proceedings*, 2006.

[3] D. Aleman, H. Romeijn, and J. Dempsey. A response surface approach to beam orientation optimization in intensity modulated radiation therapy treatment planning. Technical report, Department of Industrial Engineering, the University of Florida, 2007. to appear in IIE Transactions: Special Issue on Healthcare Applications.

[4] D. Baatar and H.W. Hamacher. New LP model for multileaf collimators in radiation therapy planning. In *Proceedings of the Operations Research Peripatetic Postgraduate Programme Conference ORP^3, Lambrecht, Germany*, pages 11–29, 2003.

[5] G. K. Bahr, J. G. Kereiakes, H. Horwitz, R. Finney, J. Galvin, and K. Goode. The method of linear programming applied to radiation treatment planning. *Radiology*, 91:686–693, 1968.

[6] G.K. Bahr, J.G. Kereiakes, H. Horwitz, R. Finney, J.M. Galvin, and K. Goode. The method of linear programming applied to radiation treatment planning. *Radiology*, 91:686–693, 1968.

[7] F. Bartolozzi et al. Operational research techniques in medical treatment and diagnosis. a review. *European Journal of Operations Research*, 121(3):435–466, 2000.

[8] C. Boender and et al. Shake-and-bake algorithms for generating uniform points on the boundary of bounded polyhedra. *Operations Research*, 39(6), 1991.

[9] N. Boland, H.W. Hamacher, and F. Lenzen. Minimizing beam-on time in cancer radiation treatment using multileaf collimators. *Networks*, 43(4):226–240, 2004.

[10] S. Bonder. Improving or support for healthcare delivery systems: Guidlines from military or experience. INFORMS Annual Conference, Denver, CO, 2004.

[11] R. Caron. Personal communications.

[12] D. Cheek, A. Holder, M. Fuss, and B. Salter. The relationship between the number of shots and the quality of gamma knife radiosurgeries. *Optimization and Engineering*, 6(4):449–462, 2004.

[13] P.S. Cho, S. Lee, R.J. Marks, S. Oh, S.G. Sutlief, and M.H. Phillips. Optimization of intensity modulated beams with volume constraints using two methods: Cost function minimization and projections onto convex sets. *Medical Physics*, 25:435–443, 1998.

[14] M. Ehrgott, A. Holder, and J. Reese. Beam selection in radiotherapy design. Technical Report 95, Trinity University Mathematics, San Antonio, TX, 2005.

[15] H.W. Hamacher and K.-H. Küfer. Inverse radiation therapy planing – A multiple objective optimization approach. *Discrete Applied Mathematics*, 118(1-2):145–161, 2002.

[16] A. Holder. Partitioning multiple objective optimal solutions with applications in radiotherapy design. Technical report, Department of Mathematics, Trinity University, San Antonio, USA, 2001.

[17] A. Holder. Radiotherapy treatment design and linear programming. In M. Brandeau, F. Sainfort, and W.P. Pierskalla, editors, *Operations Research and Health*

Care: A Handbook of Methods and Applications, chapter 29. Kluwer Academic Publishers, 2004.

[18] A. Holder and B. Salter. A tutorial on radiation oncology and optimization. In H. Greenberg, editor, *Emerging Methodologies and Applications in Operations Research*. Kluwer Academic Press, Boston, MA, 2004.

[19] E.K. Lee, T. Fox, and I. Crocker. Integer programming applied to intensity-modulated radiation therapy treatment planning. *Annals of Operations Research*, 119:165–181, 2003.

[20] G. Lim, J. Choi, and R. Mohan. Iterative solution methods for beam angle and fluence map optimization in intensity modulated radiation therapy planning. Technical report, Department of Industrial Engineering, University of Houston, Houston, Texas, 2007. to appear in the Asia-Pacific Journal of Operations Research.

[21] J. Lim, M. Ferris, D. Shepard, S. Wright, and M. Earl. An optimization framework for conformal radiation treatment planning. Technical Report Optimization Technical Report 02-10, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, 2002.

[22] S. Morrill, I. Rosen, R. Lane, and J. Belli. The influence of dose constraint point placement on optimized radiation therapy treatment planning. *International Journal of Radiation Oncology, Biology, Physics*, 19:129–141, 1990.

[23] P. Nizin and R. Mooij. An approximation of centeral-axis absorbed dose in narrow photon beams. *Medical Physics*, 24(11):1775–1780, 1997.

[24] W. Pierskalla. We have no choice - health care delivery must be improved: The key lies in the application of operations research. INFORMS Annual Conference, Denver, CO, 2004.

[25] I. Rosen, R. Lane, S. Morrill, and J. Belli. Treatment plan optimization using linear programming. *Medical Physics*, 18(2):141–152, 1991.

[26] D. Shepard, M. Ferris, G. Olivera, and T. Mackie. Optimizing the delivery of radiation therapy to cancer patients. *SIAM Review*, 41(4):721–744, 1999.