Beam Angle and Fluence Map Optimization for PARETO Multi-Objective Intensity Modulated Radiation Therapy Treatment Planning

by

Heather Champion

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

Master of Science

Department of Physics and Astronomy
University of Manitoba
Winnipeg

ABSTRACT

In this work we introduce PARETO, a multiobjective optimization tool that simultaneously optimizes beam angles and fluence patterns in intensity-modulated radiation therapy (IMRT) treatment planning using a powerful genetic algorithm. We also investigate various objective functions and compare several parameterizations for modeling beam fluence in terms of fluence map complexity, solution quality, and run efficiency. We have found that the combination of a conformity-based Planning Target Volume (PTV) objective function and a dose-volume histogram or equivalent uniform dose -based objective function for Organs-At-Risk (OARs) produced relatively uniform and conformal PTV doses, with well-spaced beams. For two patient data sets, the linear gradient and beam group fluence parameterizations produced superior solution quality using a moderate and high degree of modulation, respectively, and had comparable run times. PARETO promises to improve the accuracy and efficiency of treatment planning by fully automating the optimization and producing a database of non-dominated solutions for each patient.

ACKNOWLEDGEMENTS

First, I would like to thank my supervisors, Dr. Boyd McCurdy and Dr. Jason Fiege. Boyd is a talented Medical Physicist with a contagious enthusiasm for research. His knowledge and guidance have helped me to gain a deep appreciation for the scope of this work, and have greatly contributed to the project's success. Jason is a gifted computational scientist: Astrophysicist by day and Optimization Expert by night. His multiobjective evolutionary algorithm is the powerhouse optimizer behind the PARETO treatment planning system. As my mentor in both undergraduate and graduate work, Jason has stirred my interest in computational physics, sharpened my problem solving skills, and even provided a template for my coding style (minus the comments).

I would also like to thank the rest of the PARETO team: Dr. Peter Potrebko and Andrew Cull. Peter's Ph.D. work in beam orientation optimization provided the initial motivation for PARETO. Peter has shown that PARETO is a cutting edge technology capable of handling a variety of treatment planning problems, and has always been persistent in keeping PARETO clinically motivated. Andrew's undergraduate work helped to turn PARETO from a concept into a working prototype. All members of the PARETO team have gratefully exploited his programming proficiency. Together our team has laboured fervently (at times, literally around the clock) for abstract deadlines, conference presentations, patent applications, and our first publication. Thank you Boyd, Jason, Peter, and Andrew for a very productive and memorable couple of years.

Thanks also to my friends and colleagues at CancerCare for stimulating conversations, inspirational quotes, and their company and support. These people include Tamar Chighvinadze, Ganiyu Asuni, Troy Teo, Hongyan Sun, Mike Hebb, Peter McCowan, Jenna King, Timothy VanBeek, Hillgan Ma, Esther MacKinlay, Deanne Dombrosky, Dr. Jorge Alpuche, Dr. Krista Chytyk, Dr. Niranjan Venugopal, Dr. Eric VanUytven, Dr. Harry Ingleby, Dr. Daniel Rickey, and Dr. Ingvar Fife. I would also like to thank my mother for her encouragement and support throughout my university studies. I am also grateful for my father, who was a gifted science teacher with an inspiring passion for the truth. He bravely fought with cancer over the course of eight years.

Finally, I would like to acknowledge the generous financial support of the Natural Sciences and Engineering Research Council, the Manitoba Health Research Council, CancerCare Manitoba, and the University of Manitoba.

To my father, Geoff Champion

TABLE OF CONTENTS

SUMMARY OF CONTRIBUTIONS

This thesis consists of three chapters. I wrote the text of Chapter 1, which provides an introduction to radiotherapy treatment planning. I adapted Chapter 2 from a paper our group published in *Medical Physics* entitled “PARETO: A Novel Evolutionary Optimization Approach to Multi-Objective IMRT Planning” (*Med. Phys.* 38 (9), pp. 5217-5229, 2011). For this paper, I was personally responsible for generating and presenting all of the results. I also contributed to the early development work mentioned in the paper, including testing various formulations of the fitness functions, implementing the ‘inner loop’ fluence optimization approach, and testing many versions of the linear gradient fluence parameterization. Also, as I was developing a new fluence parameterization method (the isodose projection method) I identified the need for an optimized beam’s-eye-view margin parameter, then I implemented it and used it to generate the results of Chapter 2 and 3. I also contributed to the methodology for comparing PARETO solutions with a commercial treatment planning system. In Chapter 3, I present my investigation of various fluence parameterizations. An abbreviated version of this chapter may be published in the near future. I was largely responsible for the detailed implementation of each fluence parameterization method and fluence complexity metric, and entirely responsible for generating and presenting the results of Chapter 3. I also invented the “round robin tournament” as a scalar metric for evaluating the overall quality of a PARETO solution database.

LIST OF FIGURES

LIST OF TABLES

GLOSSARY OF TERMS

3D-CRT	Three-dimensional Conformal Radiation Therapy
BAO	Beam Angle Optimization
BEV	Beam's-Eye-View
BIV	Beam Intersection Volume
CT	Computed Tomography
CTV	Clinical Target Volume
DAO	Direct Aperture Optimization
DCT	Discrete Cosine Transform
DICOM	Digital Imaging and Communications in Medicine
DNA	Deoxyribonucleic Acid
DVH	Dose-Volume Histogram
EBR	External-Beam Radiotherapy
EUD	Equivalent Uniform Dose
FMO	Fluence Map Optimization
FSU	Functional Subunit
GA	Genetic Algorithm
GTV	Gross Tumour Volume
ICRU	International Commission on Radiation Units and Measurements
IMRT	Intensity-Modulated Radiation Therapy
LINAC	Linear Accelerator
MC	Monte Carlo
MLC	Multi-leaf Collimator

MRI	Magnetic Resonance Imaging
OAR	Organ-At-Risk
OHT	Other Healthy Tissue
PARETO	Pareto-Aware Radiotherapy Evolutionary Treatment Optimization
PB	Pencil-Beam
PRV	Planning Organ-At-Risk Volume
PTV	Planning Target Volume
ROI	Region-Of-Interest
SF	Survival Fraction
SSD	Source to Surface Distance
STD	Standard Deviation
TERMA	Total Energy Released per unit Mass
TPS	Treatment Planning System
VOI	Volume Of Interest

1 INTRODUCTION

1.1 Fundamentals of Radiation Therapy

Cancer incidence and mortality rates are on the rise due to a growing and aging Canadian population [1]. In 2011, 177,800 new cases of cancer and 75,000 deaths from cancer were estimated to occur in Canada [1]. Radiation therapy is a chief treatment modality that is used to treat approximately 50% of all cancer patients. The leading technology for delivery of external-beam radiotherapy (EBR) is the medical linear accelerator (LINAC). A LINAC produces a high-energy photon beam by accelerating electrons along a waveguide towards a target. Upon impact with the target, the electrons produce bremsstrahlung photons that are passed through a flattening filter to promote a uniform incident intensity distribution. A multi-leaf collimator (MLC) may then be used to create complex intensity distributions and apertures that shape the beam. In static delivery mode, the MLC leaves remain stationary while the beam is on. However, complex intensity patterns can be delivered efficiently via a dynamic mode, in which leaf motion occurs while the beam is on. Both the LINAC head (gantry) and patient couch can rotate 360° in planes perpendicular to each other. Often, radiation is delivered at a set of fixed gantry and patient couch angles (see Fig. 1.1). Alternatively, radiation may be delivered while the gantry is moving along an arc (arc therapy).

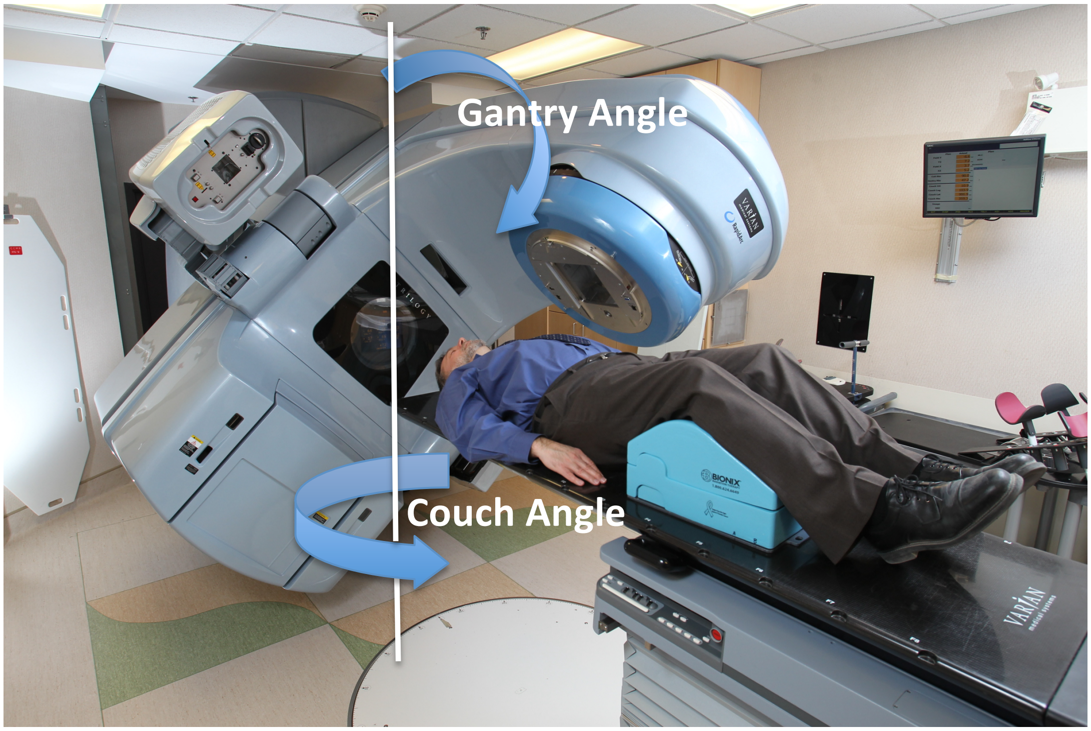

Figure 1.1: A medical LINAC used for radiotherapy. Photo courtesy of Marianne Helm.

For the energy range used in radiotherapy, the dominant interaction mechanism of radiation with matter is Compton scattering. In the Compton effect, a photon interacts with a loosely bound orbital electron and is scattered away at a lower energy. The electron is ejected and interacts with atoms and electrons along its path, gradually transferring its kinetic energy to the surrounding medium. Radiation dose represents the absorbed energy per unit mass of tissue, in units of Gray (1 Gy = 1 J/kg).

When radiation ionizes the water molecules of tissue, free radicals may form which can cause deoxyribonucleic acid (DNA) strand breaks that will lead to cell death if not repaired properly. Also, radiation can directly ionize the atoms of the DNA strand. Cancer cells are generally rapidly dividing, undifferentiated, and early responders to radiation damage [2]. Therefore, designing a schedule of radiation treatments in which the dose is delivered in small daily fractions helps to increase the damage to the tumour since cancer cells will rapidly redistribute into radiosensitive phases of the cell cycle (this is called ‘reassortment’). Furthermore, a fractionated schedule will also allow reoxygenation of hypoxic regions of the tumour, which will increase tumour damage since the presence of oxygen increases radiosensitivity. Meanwhile, healthy cells will benefit from the time between fractions as they repopulate and repair sub-lethal DNA strand breaks. At each fraction, a treatment is planned with the goals of killing tumour cells and minimizing the dose to healthy tissue.

1.2 Treatment Planning Methodology

Treatment planning is conducted by evaluating the quality of a given dose distribution with respect to goals defined for each clinical volume of interest (VOI). Clinicians use the designations of the International Commission on Radiation Units and Measurements (ICRU) to define VOIs [3, 4]. The gross tumour volume (GTV) includes the total extent of the tumour visible in any three-dimensional (3D) image, as well as lymph nodes and locations of metastatic disease. However, due to the likelihood of microscopic disease, a margin is added to the GTV to define the clinical target volume (CTV). An additional margin must also be added to account for patient motion, patient positioning uncertainties, and interfraction organ size variation. The volume of the CTV plus this margin is defined as the planning target volume (PTV). Organs-at-risk (OARs) are defined as structures whose radiosensitivity may influence treatment planning. In analogy to the PTV, a margin may be added to the OARs to accommodate patient motion and setup uncertainties, defined as the planning organ-at-risk volume (PRV).

For each treatment plan, as part of an optimization process (see Section 1.3) or for a final analysis, a dose distribution must be calculated over the irradiated volume, including the PTV, OARs, and surrounding other healthy tissue (OHT). Presently, dose calculations are performed to a high degree of accuracy with the aid of computer algorithms. In 2004, the American Association of Physicists in Medicine (AAPM) reported that uncertainties in dose calculation algorithms range from 1-5%, but they anticipated that these would quickly reduce to 1-3% with improved imaging and radiation technology [5]. However, when the total uncertainty for the complete treatment procedure is considered, 2-3% error in dose computation contributes to approximately 5% or larger total error [5]. The most accurate dose calculation algorithm is the Monte Carlo (MC) method, which simulates the individual motion of photons and electrons and records energy deposition in the patient. Thus, MC algorithms account for tissue heterogeneities directly and do not assume radiation equilibrium. Also, by simulating the motion of particles through the MLC, radiation leakage and scatter are accurately measured [6]. However, MC algorithms are currently too slow for iterative use in optimization algorithms.

Correction-based dose calculation algorithms include broad-beam algorithms and pencil beam algorithms. Broad-beam algorithms are useful in situations where the radiation intensity (fluence) is constant or smoothly varying across a beam field. First, the dose distribution for a simple homogenous water phantom is specified as a function of field size and source to surface distance (SSD) using measured data. The dose distribution for a particular plan is found by correcting this for treatment and patient-specific parameters such as surface contours and tissue heterogeneities [6]. Correction-based pencil-beam (PB) algorithms, on the other hand, may be used in situations where there is a non-uniform fluence distribution since they consider a radiation beam as a set of pencil beams with different intensities. Measured data is convolved with pre-computed energy deposition kernels (produced by MC simulation or empirical data), which accounts for the intensity distribution of the beam, the field shape, and lateral transport of secondary radiation [6]. Patient heterogeneities and anatomy are corrected for using radiological path length scaling methods. PB algorithms provide an effective compromise between accuracy and computation speed and are used routinely in treatment plan optimization (see Section 1.3).

Finally, model-based, point-interaction kernel algorithms are second only to MC algorithms in their accuracy, but are much faster to evaluate computationally. Kernel algorithms separate the effect of primary and secondary photons within the patient. Primary photon energy is described by the total energy released per unit mass (TERMA) which is the product of the mass attenuation coefficient, $\frac{\mu}{\rho}(r')$, and the energy fluence distribution $\Psi(E, r')$ at each interaction site r' in the patient:

$$T(r') = \frac{\mu}{\rho}(r') \cdot \Psi(E, r'). \qquad (1.1)$$

The variable μ represents the linear attenuation coefficient, while ρ represents the density. Secondary radiation is modeled with a pre-computed kernel $K(r \cdot \rho;\ r' \cdot \rho')$ that is in general spatially variant. Kernels are often density-scaled according to radiological path length, $\tilde{\rho} \cdot (r - r')$, where $\tilde{\rho}$ represents the average density along the path between the interaction site (r') and the dose deposition site (r) [7]. Kernels for polyenergetic spectra may be calculated by Monte Carlo methods [8], or modeled analytically [9]. Tissue heterogeneities may be accounted for by attenuating the primary radiation TERMA and applying density-scaling on the secondary kernel [5]. The total dose deposited in tissue is computed by a superposition of the TERMA with the dose kernel:

$$D(r) = \int T(r' \cdot \rho') K(r \cdot \rho;\ r' \cdot \rho')\, d^3 r'. \tag{1.2}$$

If heterogeneity corrections are omitted and the kernel is spatially invariant, the calculation reduces to a convolution such that K is merely a function of the relative position of the interaction and dose deposition sites $(r - r')$ [5]. Kernel-based superposition algorithms take longer to compute than PB algorithms, but they are more accurate in heterogeneous geometries, and are used in many treatment planning systems (TPSs) [6].

Most commonly, dose distributions are evaluated according to how well they conform to the PTV and spare OARs. The dose-volume histogram (DVH) provides a straightforward way to visualize the characteristics of a dose distribution for each VOI. A differential DVH sums the number of volume elements (voxels) within a VOI that have a dose level that falls within a given bin. However, cumulative DVHs, which represent the total sum of all voxels at or above a certain dose level, are used in clinical practice to quickly read off conventional dose statistics such as the minimum, median, and maximum dose of a VOI. A given plan is often deemed acceptable if at least 95% of the

PTV receives the prescription dose and the tolerance of OARs is not exceeded [10]. The tolerance of some normal tissues depends on the spatial arrangement of functional subunits (FSUs). FSUs are used to describe compartments of an organ that each perform part of the organ's function. If FSUs are arranged in parallel, they exhibit a graded dose response such that organ function depends on the volume of the structure irradiated. However, if FSUs are arranged in series, each one is critical to organ function, and so response depends more critically on the volume irradiated [2].

Driven by the clinical goals of sparing OARs and delivering a uniform prescription dose to the PTV, EBR technology has undergone two major developments that have each improved and changed the nature of treatment planning. First, 3D imaging technologies such as computed tomography (CT) and magnetic resonance imaging (MRI) have given rise to 3D conformal radiation therapy (3D-CRT). In 3D-CRT, the ROIs are first outlined on a 3D image and then projected onto the beam's-eye-view (BEV) planes. These projections allow beam apertures to be defined that conform the dose to the PTV and block or at least partially block the OARs [11]. The total fluence delivered at each beam angle is generally constant across the beam, though the weight of each beam may be adjusted. While 3D-CRT offers fair tumour coverage and OAR sparing using a simple geometric planning procedure, it is not guaranteed to produce a uniform dose distribution over complex-shaped PTVs [12].

Another milestone in the development of EBR technology was the advent of intensity-modulated radiation therapy (IMRT), which delivers complex, non-uniform fluence patterns with an MLC. Superior dose distributions may be achieved with IMRT. For instance, in many cases where a concave PTV wraps around an OAR of serial FSU structure, a steep increase in beam fluence is needed towards the concavity in order to produce a homogeneous dose distribution in the PTV (see Fig. 1.2) [12]. Also, there are some situations in which an intentionally inhomogeneous dose distribution may be desired, such as when an OAR lies within the expansion margin created for motion and set-up uncertainty around the CTV, or in regions of subclinical disease where moderate dose levels are sufficient to achieve high rates of control [12]. Furthermore, IMRT has the theoretical capability of compensating for the local loss of electron equilibrium due to interfaces between low and high-density media. However, since the dose calculation algorithms used in clinical practice do not yet model secondary electron transport well, IMRT is not yet used in this context [12]. Since the complexity of the IMRT planning problem is much greater than that of 3D-CRT, IMRT planning *must* be performed with the aid of an optimization algorithm (see Section 1.3 below).

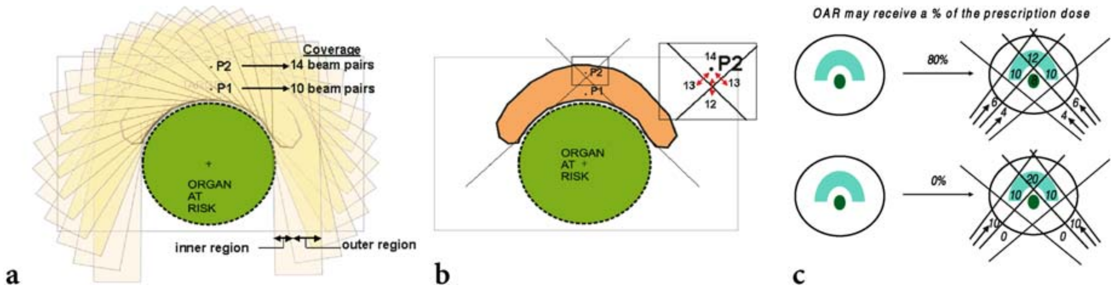

Fig. 1. (a) Irradiation of a concave PTV by many static beams that are tangential to the OAR. Let us assume that each rectangle represents a pair of opposed beams and that for each pair the variation of dose deposition with depth is negligible. Inside the PTV, point P1 is located closer to the concavity than point P2. The figure shows that P1 is exposed to fewer beams than P2 and in general, the closer a point is located nearby the concavity, the fewer the number of beams it is exposed to. If the intensity across the beam pair cross section is equal (unmodulated) then the dose to a point is proportional to the number of beam pairs that it is exposed to. Thus, the dose to P1 is 10/14th of the dose of P2. Let us divide all beam pairs in an inner (close to the concavity) and an outer region. Analysis of the figure shows that point P1 is never exposed to the outer region. Point P2 is exposed to the inner region of some beam pairs and to the outer region of other beam pairs. Thus by reducing the intensity in the outer region of the beam pairs, the dose to P2 can be lowered selectively. By repeating this procedure for pairs of points inside the PTV, we can show that the intensity inside the beam pair increases with decreasing distance to the concavity. (b) Let us consider point P2. Point P2 is embedded at the intersection of 2 beam pairs. The *insert* shows that P2 receives dose from these 2 beam pairs on top of dose from 12 beam pairs. At the beam edges inhomogeneities occur as indicated by the *arrows*. Note that the beam edges nearby the highest intensity are concerned. The *vertical arrow* crosses an inhomogeneity by exposure to n vs $n-2$ beam pairs, n being 14 in this example. The *oblique arrows* cross inhomogeneity by exposure to n vs $n-1$ beam pairs. The inhomogeneity will be roughly proportional to $n/n-a$, the value of a being dependent on the intersections of beam edges. Thus, with decreasing n, the inhomogeneity will increase. As a consequence the regional inhomogeneity inside the target will be larger closer to the concavity. (c) Example of a decreasing in-PTV inhomogeneity forced by increasing the requested dose ratio between the OAR and the concave PTV. In the upper panel, the OAR may receive 80% of the median prescription dose and the inhomogeneity is smaller than in the lower panel where the OAR may receive 0% of the prescription dose

Figure 1.2: With kind permission from Springer Science+Business Media: *Image-Guided IMRT,* "Rationale of Intensity Modulated Radiation Therapy: A Clinician's Point of View", 2006, p. 5, W. De Neve, Fig. 1.

1.3 Optimization

Traditionally, 3D-CRT plan optimization is classified as a forward planning process, where the treatment variables (such as gantry angle, wedge angle, and beam weight) are iteratively adjusted until the resulting dose distribution is deemed acceptable. On the other hand, IMRT optimization is distinguished as an inverse planning procedure. However, a true inverse planning approach would involve specifying the desired dose distribution and calculating the treatment variables that reproduce it. Due to physical constraints, it is not possible to achieve a uniform dose everywhere inside the tumour with no dose outside [13]. Therefore, the next "best" dose distribution that is also physically feasible is actually unknown to the treatment planner. Moreover, since it is not possible to completely spare all the OARs and deliver a uniform dose to the PTV, there are many different compromises that may be made (see below). Therefore, mathematical functions are formulated that describe clinical objectives for the critical structures, and an optimization engine is used to find an optimal dose distribution. In this approach, the optimization algorithm performs ‘forward planning’ by varying the parameters automatically, while the human operator practices ‘inverse planning’ by specifying the objective functions. Still, usually some of the treatment variables are iterated manually, such as the beam angles (see Fig. 2.1).

There are multiple goals involved when treating with radiotherapy. In general, it is desired to produce a homogeneous dose distribution inside the tumour and to minimize the dose to the OARs and OHT. Since many optimization algorithms require convex objective functions so as not to get trapped in local minima, each goal is typically formulated by the sum of quadratic dose deviations for a VOI. Thus, typical objective functions for underdosage and overdosage may be [14]:

$$OF_k^{(-)}(x) = \frac{1}{N_k}\sum_{i=1}^{N_k}\left[C_+(D_{min}^k - D_i^k(x))\right]^2, \tag{1.3}$$

$$OF_k^{(+)}(x) = \frac{1}{N_k}\sum_{i=1}^{N_k}\left[C_+(D_i^k(x) - D_{max}^k)\right]^2. \tag{1.4}$$

The operator C_+ ensures that only violations above or below the maximum or minimum dose, respectively, are penalized:

$$C_+(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}. \tag{1.5}$$

There are two major approaches for handling the conflicting goals of tumour coverage and OAR sparing. The first is called scalarization, where the multiple objectives are collapsed into a single objective function using weighting factors. In this case, the human operator must decide on the relative importance of each term in the objective function. For example, the PTV may require terms for dose violations above and below the maximum and minimum dose, respectively, and each OAR requires a term for overdosage [14]:

$$OF(x) = w_{PTV}^{(+)} OF_{PTV}^{(+)}(x) + w_{PTV}^{(-)} OF_{PTV}^{(-)}(x) + \sum_k w_k OF_k^{(+)}(x). \qquad (1.6)$$

The variables $w_{PTV}^{(+)}$ and $w_{PTV}^{(-)}$ represent the relative weight applied to overdosed and underdosed voxels in the PTV. Each OAR is assigned a weight w_k that represents the priority of reducing overdosed voxels. Often the human operator will iteratively vary the weighting factors in order to arrive at their preferred plan. In the process, the planner is rejecting solutions that represent different dose priorities. An alternative approach is for the decision maker to specify their preferences after performing an optimization that varies the weighting factors automatically. This is a technique commonly employed in multiobjective optimization, resulting in a set of optimal solutions that represent various feasible tradeoffs.

In multiobjective optimization, the goal is to sample the multidimensional tradeoff surface with Pareto optimal solutions. A Pareto optimal solution is one where no one objective can be improved without degrading at least one other. For a set of *n* objective functions, vectors $\mathbf{u} = (u_1, \dots, u_n)$ and $\mathbf{v} = (v_1, \dots, v_n)$ each consist of the objective function values of two different solutions. $\mathbf{u}$ is said to be inferior to $\mathbf{v}$ if and only if $\mathbf{v}$ is partially less than $\mathbf{u}$ (i.e. $\forall i = 1, \ldots, n, v_i \leq u_i \ \wedge \ \exists i = 1, \ldots n: v_i < u_i$). $\mathbf{u}$ is said to be superior to $\mathbf{v}$ if and only if $\mathbf{v}$ is inferior to $\mathbf{u}$ [15]. A solution is non-dominated within a set of solutions if it is neither inferior nor superior to any other known solution. The Pareto optimal set is comprised of all possible non-dominated solutions. It is always possible to construct a non-dominated set from a list of previously explored solutions. However, the Pareto-optimal set should be considered a mathematically ideal set of solutions that can only be found by a complete exploration of all possible solutions, which is not feasible for most problems, and certainly not the one presented in this thesis. Therefore, in this thesis we find non-dominated solutions lying on a multidimensional tradeoff surface, with the hope that they approximately sample the Pareto set. This is well-founded for many test problems, where the Pareto set is known analytically, but cannot be shown to be rigorously true for the problem of radiotherapy treatment optimization.

There are two main approaches used in radiotherapy treatment optimization for discovering a set of non-dominated solutions. As mentioned above, a weighting vector that represents the priorities of each objective function may be varied systematically [16]. However, this approach cannot discover optimal solutions on non-convex regions of the tradeoff surface [17]. Another approach is the epsilon constraint method, in which each objective function or a weighted sum of objective functions is constrained during optimization [17–19]. By systematically varying the constraints, a set of non-dominated solutions is obtained. With the epsilon constraint method, non-convex regions of the tradeoff surface may be discovered. However, as discussed in Chapter 2, this approach does not scale well to problems with many objectives. Both of these methods perform a separate optimization to obtain each non-dominated solution. A single-objective optimization engine is required, which may be deterministic (e.g. gradient descent, Nelder-Mead simplex) or stochastic (e.g. simulated annealing, genetic algorithm (GA)). Deterministic approaches have the advantage of quickly finding a single optimal solution, but are only used when the objective function is convex so as to avoid getting trapped in local minima. Stochastic methods are more robust, but require a much greater computation time.

A multiobjective GA, on the other hand, uses a distinctly different approach than populating the tradeoff surface one point at a time. A GA uses the principles of biological evolution to evolve the fitness (objective function values) of a population of trial solutions over several generations (iterations). Parameters are encoded on 'genes' using binary, integer, or real number encoding. Usually, fitness is assigned to each solution (parameter set) based on a ranking method. For example, a vector comprised of the objective function values of each solution may be compared to every other solution present in the population. Points that are non-dominated are assigned a rank of one and subsequently removed from consideration. In the remaining set, non-dominated points are assigned a rank of two, and the process is repeated until all points are assigned a rank [15]. The approach used by the GA presented in this thesis is to rank each solution by the number of solutions that dominate it (see Section 2.2.1). A selection operator is used at each generation to decide which solutions will propagate to the next generation. Often a selection probability is assigned to a solution based on its fitness value, and a roulette wheel approach is used to determine whether the solution survives [17]. Another method is to compare solutions in a tournament, randomly selecting solutions to compete for survival in the next generation. When there is a tie between solutions (both are non-dominated), the solution with fewer neighbours (as determined by a niching radius in objective function space) is taken (see Section 2.2.1). Crossover operators are used to introduce new individuals into the population. In a crossover event, the genetic strings of two fit individuals are combined to create two offspring that are intermediate between the parent models. The number of crossover events and the crossover sites are determined randomly, but may be adapted dynamically in order to preserve fit individuals near the end of a run and to keep well-linked parameters together in offspring solutions. Also, GAs use mutation events to avoid getting trapped in local minima or losing potentially valuable genetic material prematurely [17]. Machine-learning may be used to automatically adapt control parameters including those related to crossover events and mutation scale as the optimization progresses (see Section 2.2.1).

Typically, IMRT treatment plan optimization is performed on parameters that represent the fluence of each beam. A comprehensive overview of common fluence parameterizations can be found in Section 3.1. Most often, beam angles are not included as parameters in the optimization, and human operators must manually adjust beam orientations until a desired solution is achieved. This is due to the fact that beam angle optimization (BAO) is known to be a difficult, non-convex problem [20]. However, if a stochastic optimizer such as a GA is used, local minima may be avoided. Most BAO algorithms use a hybrid approach where beam angles are optimized stochastically and beam fluence is handled in an 'inner loop' with faster gradient-based methods [21–25]. Some algorithms employ *a priori* knowledge of promising beam angles to inform the search [25–27]. However, unlike 3D-CRT, "good" beam angles are non-intuitive since fluence modulation can effectively spare the OARs at any beam angle. To identify promising beam angles, these algorithms often use score functions that divide each beam into a grid of beamlets, and rate the beamlets according to the maximum dose that they may deliver to the PTV without violating the tolerances of the OARs.

There is some debate as to whether BAO offers a significant improvement in solution quality. Results from early studies suggested that when using a large number (≥ 5 or 6) of equiangular-spaced beams, the plan is relatively insensitive to the specific choice of beam angles [20, 24, 28]. However, using a large number of beams may increase the volume of normal tissue exposed to low doses [29, 30]. Also, a large number of beams may increase the delivery time and therefore the probability of patient motion, and make quality assurance more cumbersome [31]. Furthermore, more recent work has shown that using only 7 or 9 beams is sufficient for achieving high quality plans [32, 33]. Thus, finding optimal beam configurations for the smallest number of beams possible is useful. Finally, the benefit of BAO may also depend on the specific treatment site [14]. Superior solution quality has been achieved for some head and neck cases with BAO [23].

Since BAO is coupled with fluence map optimization (FMO), it entails an enormous search space that may take several hours of optimization time. In conjunction with 3D-CRT, some have attempted to reduce the computation time by using geometry-based objective functions [34–37]. Much less work has been done to develop geometry-based objective functions for IMRT. In that pursuit, Potrebko *et al*. argued that for *N* incident beams, the volume of maximum dose in an OAR occurs in the *N* beam intersection volume (BIV) in that OAR [38]. Thus, instead of using dose distributions, the authors used multiple BIV components in OARs to optimize beam orientations [39]. However, this work did not use a truly multiobjective approach, and so solutions were often found with similar beam angles that corresponded to the optimal sparing of certain OARs. Therefore, this provided the rationale for early versions of a novel multiobjective BAO algorithm called PARETO (Pareto-Aware Radiotherapy Evolutionary Treatment Optimization).

PARETO uses a sophisticated multiobjective GA called Ferret to search the enormous parameter space involved in BAO and FMO and to discover a set of non-dominated solutions automatically. Ferret has been shown to perform well on massive-scale, non-convex problems [40–42]. Nonetheless, as described in Section 3.2.2, various methods have been developed that reduce the total number of parameters required to model beam fluence in order to enable a global, monolithic optimization of beam orientations and beam fluence. PARETO employs a superposition/convolution dose calculation approach in order to accurately model trial solutions. Thus, with PARETO and Ferret, the quality and efficiency of treatment planning may be directly improved.

1.4 Thesis Overview

This thesis consists of two major sections. Chapter 2 provides an overview of PARETO and Ferret, and a validation of PARETO with a commercial TPS. Results are compared for different formulations of the objective functions using a simple cylindrical phantom. The feasibility of beam number optimization is also investigated with a beam-merging algorithm. Chapter 3 presents an investigation into several different parameterizations of beam fluence using a paraspinal tumour phantom and two patient data sets. For each parameterization, the effect of increasing the number of parameters is investigated in terms of fluence complexity and solution quality. The parameterizations are then compared in terms of run time and solution quality. In summary, the goals of this work are to demonstrate PARETO's powerful ability to simultaneously optimize beam angles and beam fluence in a multiobjective framework, and to devise an efficient parameterization of beam fluence that produces sophisticated fluence patterns.

1.5 References

[1] Canadian Cancer Society's Steering Committee on Cancer Statistics, "Canadian Cancer Statistics 2011," Canadian Cancer Society, Toronto, ON, 2011.

[2] E. J. Hall, Radiobiology for the Radiologist, 5th ed. Philadelphia: Lippincott Williams & Wilkins, 2000.

[3] International Commission on Radiation Units and Measurements, "Prescribing, Recording, and Reporting Photon Beam Therapy." ICRU Report 50, Bethesda, MD, 1993.

[4] International Commission on Radiation Units and Measurements, “Prescribing, Recording and Reporting Photon Beam Therapy (supplement to ICRU report 50).” ICRU Report 62, Bethesda, MD, 1999.

[5] American Association of Physicists in Medicine, “Tissue inhomogeneity corrections for megavoltage photon beams.” AAPM Report 85, College Park, MD, 2004.

[6] J. V. Siebers, “Dose Calculations for IMRT,” in *Image-Guided IMRT*, T. Bortfeld, R. Schmidt-Ullrich, W. Neve, and D. E. Wazer, Eds. Berlin/Heidelberg: Springer-Verlag, 2006.

[7] J. J. Battista and M. B. Sharpe, “True three-dimensional dose computations for megavoltage x-ray therapy: a role for the superposition principle,” *Eng. Sci. Med.*, 15 (4), pp. 159-78, 1992.

[8] T. R. Mackie, a F. Bielajew, D. W. Rogers, and J. J. Battista, “Generation of photon energy deposition kernels using the EGS Monte Carlo code,” *Phys. Med. Biol.*, 33 (1), pp. 1-20, 1988.

[9] A. Ahnesjö, “Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media,” *Med. Phys.*, 16 (4), pp. 577-592, 1989.

[10] M. Goitein, “Designing a Treatment Plan,” in *Radiation Oncology: A Physicist’s Eye View*, New York, NY: Springer Science+Business Media, 2008, pp. 111-137.

[11] M. Goitein, “Planning Manually,” in *Radiation Oncology: A Physicist’s Eye View*, New York, NY: Springer Science+Business Media, 2008, pp. 157-175.

[12] W. De Neve, “Rationale of Intensity Modulated Radiation Therapy: A Clinician’s Point of View,” in *Image-Guided IMRT*, T. Bortfeld, R. Schmidt-Ullrich, W. Neve, and D. E. Wazer, Eds. Berlin/Heidelberg: Springer-Verlag, 2006, pp. 3-9.

[13] M. Goitein, “IMRT and ‘Optimization’,” in *Radiation Oncology: A Physicist’s Eye View*, New York, NY: Springer Science+Business Media, 2008, pp. 177-210.

[14] U. Oelfke, S. Nill, and J. J. Wilkens, “Physical Optimization,” in *Image-Guided IMRT*, T. Bortfeld, R. Schmidt-Ullrich, W. Neve, and D. E. Wazer, Eds. Berlin/Heidelberg: Springer-Verlag, 2006.

[15] C. M. Fonseca and P. J. Fleming, “Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion, and Generalization,” in *Genetic Algorithms: Proceedings of the Fifth International Conference*, 1993.

[16] D. L. Craft, T. F. Halabi, H. a. Shih, and T. R. Bortfeld, “Approximating convex Pareto surfaces in multiobjective radiotherapy planning,” *Phys. Med. Biol.*, 33 (9), p. 3399, 2006.

[17] R. T. Marler and J. S. Arora, “Survey of multi-objective optimization methods for engineering,” *Struct. Multidisc. Optim.*, 26 (6), pp. 369-395, 2004.

[18] A. L. Hoffmann, A. Y. D. Siem, D. den Hertog, J. H. a M. Kaanders, and H. Huizenga, “Derivative-free generation and interpolation of convex Pareto optimal IMRT plans.,” *Phys. Med. Biol.*, 51 (24), pp. 6349-69, 2006.

[19] S. Breedveld, P. R. M. Storchi, and B. J. M. Heijmen, “The equivalence of multi-criteria methods for radiotherapy plan optimization.,” *Physics in medicine and biology*, 54 (23), pp. 7199-209, 2009.

[20] T. Bortfeld and W. Schlegel, “Optimization of beam orientations in radiation therapy: some theoretical considerations,” *Phys. Med. Biol.*, 38, pp. 291-304, 1993.

[21] D. Djajaputra, Q. Wu, Y. Wu, and R. Mohan, “Algorithm and performance of a clinical IMRT beam-angle optimization system,” *Phys. Med. Biol.*, 48 (19), pp. 3191-212, 2003.

[22] E. Schreibmann, M. Lahanas, L. Xing, and D. Baltas, “Multiobjective evolutionary optimization of the number of beams, their orientations and weights for intensity-modulated radiation therapy,” *Phys. Med. Biol.*, 49 (5), pp. 747-770, 2004.

[23] A. Pugachev et al., “Role of beam orientation optimization in intensity-modulated radiation therapy,” *Int. J. Radiat. Oncol. Biol. Phys.*, 50 (2), pp. 551-60, 2001.

[24] J. Stein et al., “Number and orientations of beams in intensity-modulated radiation treatments,” *Med. Phys.*, 24 (2), pp. 149-60, 1997.

[25] C. G. Rowbottom, C. M. Nutting, and S. Webb, “Beam-orientation optimization of intensity-modulated radiotherapy: clinical application to parotid gland tumours,” *Radiother. Oncol.*, 59 (2), pp. 169-77, 2001.

[26] A. Pugachev and L. Xing, “Incorporating prior knowledge into beam orientation optimization in IMRT,” *Int. J. Radiat. Oncol. Biol. Phys.*, 54 (5), pp. 1565-74, 2002.

[27] E. Schreibmann and L. Xing, “Dose-volume based ranking of incident beam direction and its utility in facilitating IMRT beam placement,” *Int. J. Radiat. Oncol. Biol. Phys.*, 63 (2), pp. 584-93, 2005.

[28] S. Soderstrom and A. Brahme, “Which is the most suitable number of photon beam portals in coplanar radiation therapy?,” *Int. J. Radiat. Oncol. Biol. Phys.*, 33 (1), pp. 151-59, 1995.

[29] H. H. Liu, M. Jauregui, X. Zhang, X. Wang, L. Dong, and R. Mohan, “Beam angle optimization and reduction for intensity-modulated radiation therapy of non-small-cell lung cancers,” *Int. J. Radiat. Oncol. Biol. Phys.*, 65 (2), pp. 561-72, 2006.

[30] X. Wang, X. Zhang, L. Dong, H. Liu, Q. Wu, and R. Mohan, “Development of methods for beam angle optimization for IMRT using an accelerated exhaustive search strategy,” *Int. J. Radiat. Oncol. Biol. Phys.*, 60 (4), pp. 1325-37, 2004.

[31] S. Kim, H. C. Akpati, J. G. Li, C. R. Liu, R. J. Amdur, and J. R. Palta, “An immobilization system for claustrophobic patients in head-and-neck intensity-modulated radiation therapy,” *Int. J. Radiat. Oncol. Biol. Phys.*, 59 (5), pp. 1531-9, 2004.

[32] B. van Asselen, H. Dehnad, C. H. J. Terhaard, J. J. W. Lagendijk, C. P. J. Raaijmakers, “Segmental IMRT for oropharyngeal cancer in a clinical setting,” *Radiother. Oncol.*, 69 (3), pp. 259-66, 2003.

[33] X. R. Zhu, C. J. Schultz, and M. T. Gillin, “Planning quality and delivery efficiency of sMLC delivered IMRT treatment of oropharyngeal cancers evaluated by RTOG H-0022 dosimetric criteria.,” *Journal of applied clinical medical physics / American College of Medical Physics*, 5 (4), pp. 80-95, 2004.

[34] G. A. Ezzell, “Genetic and geometric optimization of three-dimensional radiation therapy treatment planning,” *Med. Phys.*, 23 (3), pp. 293-305, 1996.

[35] O. C. Haas, K. J. Burnham, and J. a Mills, “Optimization of beam orientation in radiotherapy using planar geometry,” *Phys. Med. Biol.*, 43 (8), pp. 2179-93, 1998.

[36] J. Meyer, S. M. Hummel, P. S. Cho, M. M. Austin-Seymour, and M. H. Phillips, “Automatic selection of non-coplanar beam directions for three-dimensional conformal radiotherapy,” *Br. J. Radiol.*, 78 (928), pp. 316-27, 2005.

[37] E. Schreibmann, M. Lahanas, R. Uricchio, K. Theodorou, C. Kappas, and D. Baltas, “A geometry based optimization algorithm for conformal external beam radiotherapy,” *Phys. Med. Biol.*, 48 (12), pp. 1825-41, 2003.

[38] P. S. Potrebko, B. M. C. McCurdy, J. B. Butler, A. S. El-Gubtan, and Z. Nugent, “A simple geometric algorithm to predict optimal starting gantry angles using equiangular-spaced beams for intensity modulated radiation therapy of prostate cancer,” *Med. Phys.*, 34 (10), pp. 3951-61, 2007.

[39] P. Potrebko, “Geometry-based beam orientation optimization for intensity modulated radiation therapy,” University of Manitoba, 2008.

[40] J. D. Fiege, D. Johnstone, R. O. Redman, and P. A. Feldman, “A genetic algorithm – based exploration of three filament models: a case for the magnetic support of the G11.11-0.12 infrared-dark cloud,” *Ap. J.*, 616 (2), pp. 925-942, 2004.

[41] J. D. Fiege, “Computational intelligence techniques for submillimetre polarization modeling,” in *ASP Conf. Ser. 343, Astronomical Polarimetry: Current Status and Future Directions*, 2005, pp. 171-175.

[42] A. Rogers and J. D. Fiege, “Gravitational lens modeling with genetic algorithms and particle swarm optimizers,” *Ap. J.*, 727 (2), pp. 80-98, 2011.

2 PARETO: A NOVEL EVOLUTIONARY OPTIMIZATION APPROACH TO MULTIOBJECTIVE IMRT PLANNING

This chapter is adapted from a manuscript entitled “PARETO: A Novel Evolutionary Optimization Approach to Multi-Objective IMRT Planning” by Jason Fiege, Boyd McCurdy, Peter Potrebko, Heather Champion*, and Andrew Cull, published in *Medical Physics* 38 (9), pp. 5217-5229, 2011.

*This author contributed to this chapter by generating all the data and figures presented. This author also helped to test early and current formulations of the fitness functions, the ‘inner loop’ fluence optimization approach, and many versions of the linear gradient fluence parameterization. This author conceived and implemented the optimized beam’s-eye-view margin parameter. This author also contributed to the methodology for the comparison of PARETO solutions with a commercial treatment planning system and performed the comparisons presented in this chapter. Finally, the author helped to edit the text of the manuscript.

2.1 Introduction

The objectives of intensity-modulated radiation therapy (IMRT) planning are to deliver a uniform prescribed dose of radiation to one or more planning target volumes (PTVs), while also minimizing the dose to each organ-at-risk (OAR). These clinical objectives are invariably in conflict, often requiring compromises when selecting the best treatment plan for a particular patient. Most commercially available approaches to optimizing IMRT plans employ a single objective fluence optimizer and simplify the problem into a single objective function through the use of relative weighting factors applied to the individual dose-objective terms [1, 2]. In clinical practice, human operators (i.e. treatment planners or dosimetrists) commonly optimize beam orientations manually, add or subtract dosimetric objectives, and vary the weights applied to the dosimetric objectives in a time-consuming, trial-and-error process, attempting to find some acceptable compromise (see Fig. 2.1). This simplification ignores the true multiple objective nature of the problem and introduces bias by the choice of weights.

In a multiobjective IMRT treatment planning approach, one would like to simultaneously discover many solutions that sample the Pareto front, which represents the optimal set of feasible compromises between competing objectives defined by the property that it is not possible to improve any objective without degrading at least one other objective. This paper proposes a novel method that uses a multiobjective genetic algorithm (GA) to solve the multiobjective problem without combining weights. Although our heuristic optimization method cannot guarantee strict mathematical Pareto-optimality, we find solutions of high quality in practice, which are very consistent between runs. (In this work, a 'run' refers to an optimization of the objective functions using the GA, and 'run time' refers to the time elapsed during the optimization.) Moreover, our approach requires no human interaction during the optimization process and scales well to problems with many objectives, which may limit the traditional epsilon-constraint method as discussed below. Thus, our method offers a practical alternative to manual beam angle optimization and automated methods that rely on optimizing a single objective function with appropriate constraints. Figure 2.1 compares a standard single-objective IMRT treatment planning approach to the multiobjective method used in this work.

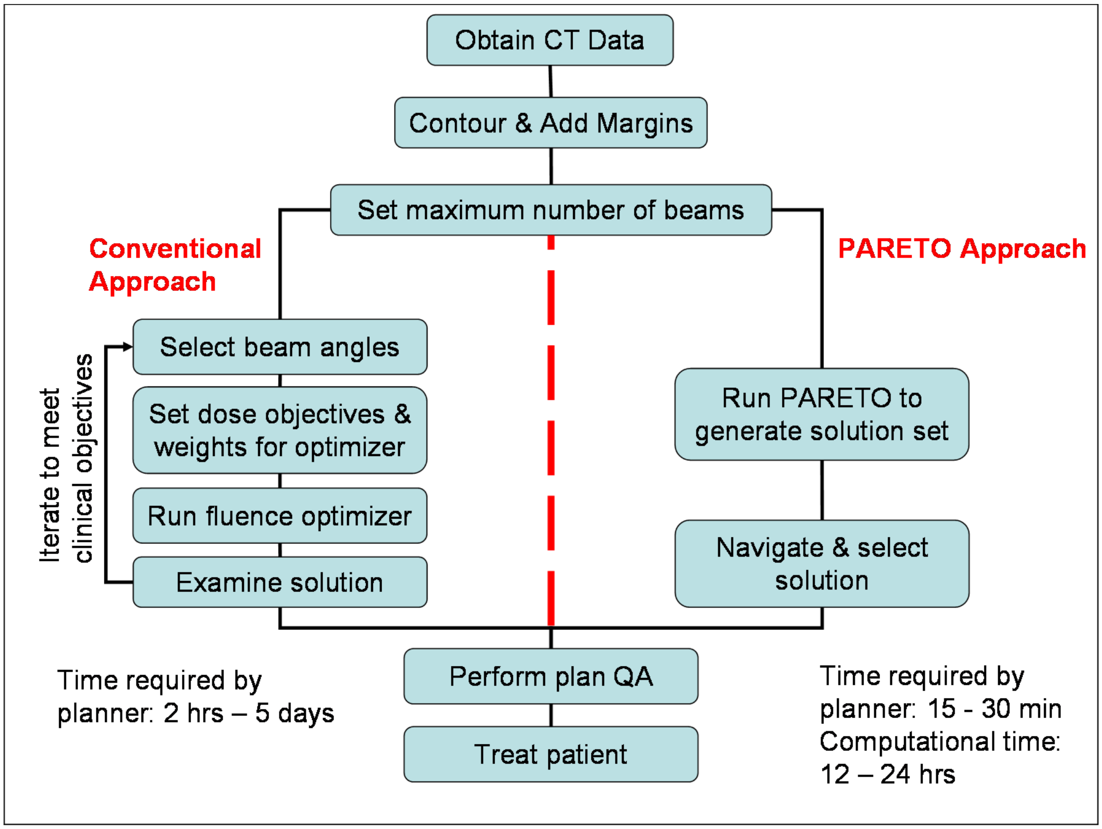
Obtain CT Data
Contour & Add Margins
Set maximum number of beams
Conventional Approach
PARETO Approach
Iterate to meet clinical objectives
Select beam angles
Set dose objectives & weights for optimizer
Run fluence optimizer
Examine solution
Run PARETO to generate solution set
Navigate & select solution
Perform plan QA
Treat patient
Time required by planner: 2 hrs – 5 days
Time required by planner: 15 - 30 min
Computational time: 12 – 24 hrs

Figure 2.1: PARETO transforms tedious manual optimization into a rapid visual search through a database of pre-optimized solutions.

A common approach to multiobjective optimization is to optimize a single-objective function constructed from the weighted sum of multiple objectives. Regions of the Pareto front can be mapped by repeatedly optimizing the weighted function while varying the weight vector between optimizations. However, it has been shown that this approach (without the inclusion of additional constraints) cannot discover optimal solutions on a non-convex region of the Pareto front [3]. Both convex and non-convex regions can be discovered by the epsilon constraint method, which optimizes a single objective (or weighted sum of objectives) from a multiobjective problem, while constraining objectives. In this approach, one minimizes a scalar weighted function $\Phi(\mathbf{f}(\mathbf{x})) = \Phi(f_1(\mathbf{x}), ..., f_m(\mathbf{x}))$ of the *m* objectives f_i specified over *n* parameters $\mathbf{x} = \{x_1, ..., x_n\}$ bounded by $x_{j,\min} < x_j < x_{j,\max}, \forall j \in \{1,...,n\}$ and subject to constraints $f_i(\mathbf{x}) < \varepsilon_i, \forall i \in \{1,...,m\}$, where ε_i is the constraint on f_i. However, mapping the Pareto front with such a method requires iterative stepping through a grid of constraints while a separate constrained optimization is performed at each step. This may be inefficient for problems with many objectives, and the step size in ε_{i} must be chosen carefully to achieve the desired resolution of the Pareto front, whose structure is not known *a priori*.

The need for non-convex solvers becomes an important issue when beam angles (or arc-orientation) are optimized simultaneously with fluence parameters. Beam fluence optimization is a convex problem for some commonly used objective functions [4], while the inclusion of beam angles into the optimization process is known to be a non-convex problem [5]. We use a sophisticated multiobjective general-purpose GA called Ferret in this work, which is designed to solve both convex and non-convex problems, as illustrated by several examples included with the software and described in its user's guide [6]. Ferret evolves a population of trial solutions over many generations, while maintaining an internal Pareto ranking of solutions. The set of non-dominated solutions is saved to disk each generation, where a non-dominated solution is one where no other solution exists within the current population that is superior in all objectives. These solutions are reloaded at the end of the run and compared against each other to construct a final set of globally non-dominated solutions, which approximates the Pareto surface for many test problems.

The concept of multiobjective optimization for radiotherapy treatment planning has been an active area of research [7-23]. Various approaches including simulated annealing optimization [7], multiobjective rotational therapy optimization [15, 19], combining optimization sequences in real-time [14], convex multiobjective fluence map optimization [10-13, 16-18, 22, 23], beam profile and voxel weight multiobjective optimization subject to soft and hard constraints [8], and conjugate gradient multiobjective optimization [9] have been studied. Schreibmann *et al.* [20, 21] investigated multiobjective optimization of the number of beams, their orientations, and intensity profiles. Their approach used a GA to perform an outer loop optimization over beam angles, while optimizing beam profiles for each set of beam angles using a gradient-based algorithm in an inner loop. We propose a simpler approach that treats beam angles and fluence parameters as part of a single, monolithic, multiobjective optimization problem, where all parameters are treated equally. This approach is efficient because it does not require division of the problem into inner and outer optimizations, but a powerful optimizer is required to avoid local minima in the enormous search space.

In this paper, we introduce a new software package named PARETO (Pareto-Aware Radiotherapy Evolutionary Treatment Optimization), which uses the Ferret GA to optimize this difficult multiobjective problem. PARETO and Ferret operate at arm's length; PARETO is entirely responsible for the specification of the radiotherapy problem and the evaluation of the fitness functions, while Ferret functions as a black box optimizer whose role is to choose "good" parameters for PARETO to evaluate without any internal knowledge of the radiotherapy problem. We describe the evolutionary optimizer used to solve the monolithic radiation treatment planning problem in Section 2.2.1 and provide an overview of PARETO in Section 2.2.2. We describe our phantoms in Section 2.2.3 and discuss our methodology for comparing to a commercial treatment planning system in Section 2.2.4 and show an actual comparison in Section 2.3.1. In Section 2.3.2, we compare several different choices of fitness functions (i.e. objective functions) currently implemented within PARETO for a simple phantom geometry and coplanar beam arrangements and explore the usefulness of PARETO for selecting the best number of beams used for treatment. We discuss our results in Section 2.4, where we also indicate PARETO's current limitations and plot a course toward a feasible tool for clinical use. Final conclusions are provided in Section 2.5.

2.2 Methods and Materials

2.2.1 The Ferret Genetic Algorithm

The solution to a multiobjective optimization problem is given by the set of points residing on a multi-dimensional Pareto front (or trade-off surface) of dimensionality $N_{obj} - 1$ for a problem with N_{obj} non-degenerate objective functions. For example, the Pareto set of a dual objective problem usually forms a one-parameter curve in the objective plane and parameter space, which may be continuous or piecewise continuous. Likewise, the solution set to a problem defined by three objectives normally spans a two-parameter curved surface. Multi-objective optimizers strive to sample the Pareto front with solutions, which represent the best compromises possible between competing objectives.

GAs are an important class of algorithms for global optimization that work in analogy to biological evolution. A basic GA encodes parameters on “genes”, which are expressed as a model (or individual) and evaluated by a fitness function. A population of such models evolves over multiple generations by applying mutation, crossover, and selection operators [24], which are designed to explore the parameter space and gradually improve the fitness of the population. The role of mutation is to apply occasional random perturbations to individuals, which helps the algorithm to explore new regions of the parameter space. Crossover combines two individuals to produce offspring that are intermediate between the parent models. The selection operator is an information filter based on the Darwinian notion of survival of the fittest, which preferentially selects fit individuals to propagate to the next generation, while destroying weaker individuals. Various types of selection operators are possible, but tournament selection has the advantage that it is insensitive to the scaling of the fitness functions [25].

The Ferret GA is a parallel multiobjective GA that has been under development since 2002 [6]. Ferret provides a thorough exploration of PARETO's parameter space and the ability to map trade-off surfaces between its multiple objective functions, which allow the user to understand the compromises that must be made during treatment planning. The interface provides an indication of convergence by graphically monitoring the number of solutions present in the per-generation non-dominated set, the median rank of the population, and the minimum value achieved for each objective. We ran the code far past apparent convergence for all results presented in this work, which has the effect of increasing the number of solutions in the optimal set. As discussed briefly in the previous section, this optimizer constructs a final database of non-dominated solutions at the end of a run by reloading all saved solutions that were non-dominated within their own generation, comparing them against each other iteratively, and rejecting any solution that is dominated by a solution from another generation. The database is guaranteed to be a non-dominated set constructed from thousands of trial solutions explored during the run. This non-dominated set is not proven mathematically to coincide with the Pareto surface, but Ferret has been shown to accurately map the Pareto surface on several test problems [6]. Moreover, we find solutions in practice that compare favorably with those found from commercial treatment planning systems. Results are almost perfectly consistent between runs, which is encouraging because such consistency would be unlikely for a stochastic optimizer struggling to find the Pareto surface and becoming trapped by false solutions far from Pareto-optimality.

Ferret uses a niching algorithm similar to the one discussed by Fonseca *et al.* [26] to spread PARETO's solutions approximately uniformly over the final optimal set. A niching algorithm measures the Euclidean distance, in either parameter or objective space, between trial solutions at each generation and calculates a niche count for each solution, which is given by the number of near neighbors within a pre-defined niching radius. The Pareto rank of a solution is given by $R = 1 + N_{\text{sup}}$, where N_{sup} is the number of superior (dominating) solutions present, such that $R = 1$ for the non-dominated set. Ties in rank are common in a multiobjective GA when two solutions compete in tournament selection. When ties occur, the diversity of the population is improved by keeping the solution with the lower niche count, since it resides in a less well-represented (and probably less well-explored) region of the parameter or objective space. PARETO was configured to perform niching in objective space for all results presented in this paper.

PARETO makes extensive use of other features in Ferret that further extend the basic GA paradigm for parameter search and global optimization [24]. For instance, this GA contains a machine-learning algorithm that monitors optimization progress to automatically adapt the mutation scale, the size scale of crossover events, and several other important control parameters in order to improve the search in response to the fitness landscape. Auto-adaptation provides an extra layer of robustness to PARETO by making the search less sensitive to the GA's internal control parameters (strategy parameters). A second machine learning technique, known as linkage-learning, studies the non-linear structure of the problem during optimization and searches for opportunities to break large, nearly intractable problems into several smaller problems that are easier to solve. During each run, the GA monitors function evaluations for a particular type of non-linear behaviour, where varying two parameters A and B separately degrades the fitness, but varying them together results in an improvement. When this type of non-linearity is detected, parameters A and B are considered to be linked, so that they are traded together preferentially during crossovers. An extension of this technique allows larger groups of linked parameters to be discovered, as discussed in [6]. A high-performing linkage group represents a distinct component of a solution specified by a good set of parameters. Thus, it is beneficial to exchange each known linkage group as a unit during crossover to preserve high-performing groups and allow them to replicate throughout the population. The current implementation of PARETO enables Ferret's linkage-learning feature on parameters representing beam angles, where it appears to improve convergence. Linkage-learning is not currently enabled on the much larger number of fluence parameters because this would be computationally expensive, and we have not found any evidence for improved convergence.

GAs are well suited for parallel computing because each individual in the population represents a single parameter set, which can be evaluated independently of other parameter sets. PARETO takes advantage of Ferret's internal parallelization, which dramatically speeds up runs performed on multi-CPU computers and inexpensive clusters. Ferret is part of the Qubist Global Optimization Toolbox, which is distributed by nQube Technical Computing Corporation (Winnipeg, Canada). Older versions of the code [27, 28] and the current version [29, 30] have been used for various problems in astronomy and astronomical instrumentation.

2.2.2 An Overview of PARETO

Figure 2.2 shows the essential components of the Ferret GA and the procedure for specifying a radiotherapy treatment plan with PARETO. For each patient, a DICOM (Digital Imaging and Communications in Medicine) Computed Tomography (CT) data set is converted into Matlab format and preprocessed to define regions of interest (ROIs). The essential function of PARETO is to provide a dose model for a given parameter set that may be used to evaluate the fitness functions. However, previous work leading to PARETO focused on geometric rather than dose-based optimization methods that minimized the volume-weighted average number of beams that intersected OARs [31]. Geometry-based fitness functions are attractive because they can be computed much more rapidly than dose-based methods. Rapid evaluation is especially important for our approach, since Ferret typically requires 10^4 to 10^5 evaluations of the fitness functions to optimize problems involving more than about five parameters, with the required number of evaluations increasing only very slowly with the size of the problem [6]. Nonetheless, purely geometric methods were abandoned because they did not produce consistently better plans than manual approaches at some treatment sites.

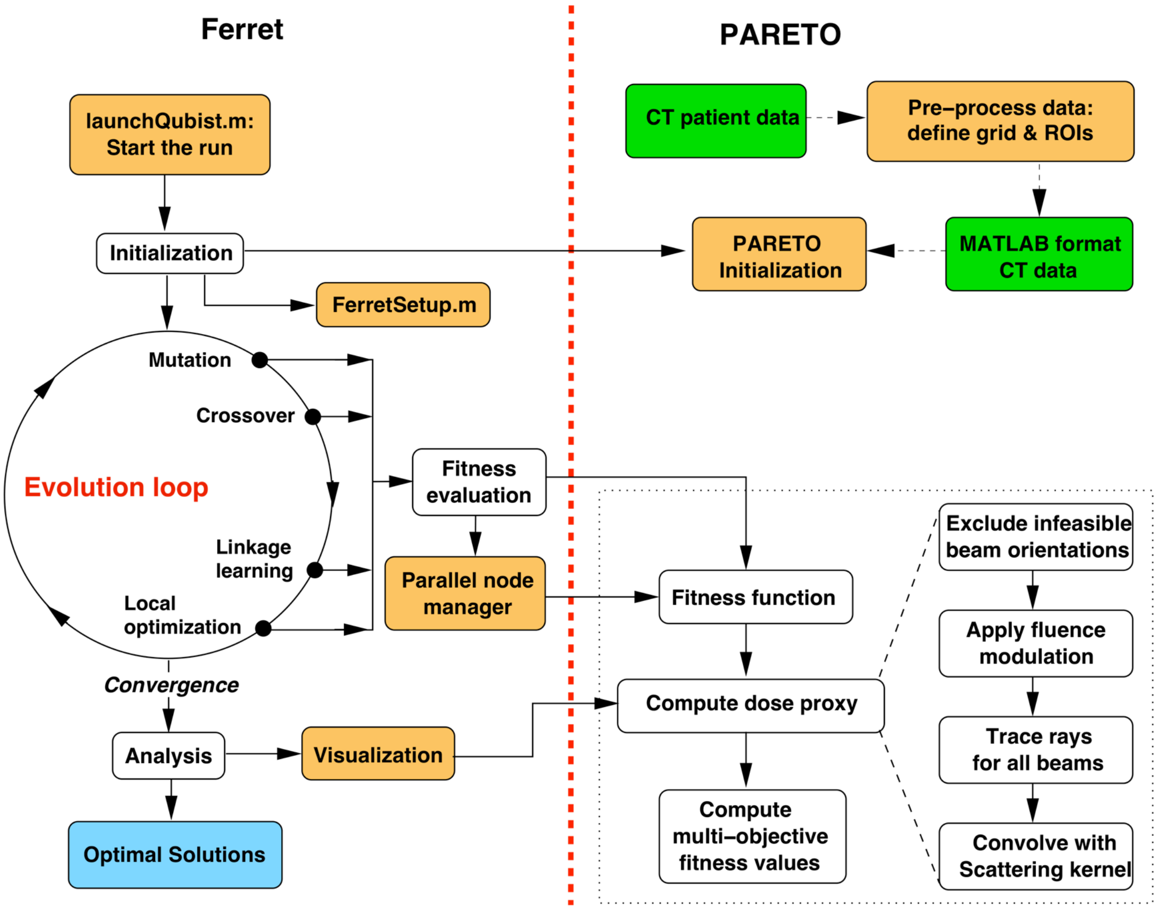

Figure 2.2: Flow chart showing Ferret and PARETO components. Ferret functions as a 'black box' optimizer for the treatment plans modeled by PARETO.

The current version of PARETO utilizes ray-tracing methods to calculate the primary dose, including divergence, attenuation, and inverse-square law effects, which is then convolved with a scattering kernel to simulate patient scatter [32]. Rays are assumed to be conformal to the projection of the PTV in the beam's-eye-view (BEV) plane, and are modulated by a parameterized fluence pattern at each beam angle. We initially attempted to optimize fluence patterns by solving an inner loop fluence optimization problem for each pair of beam angles (θ, ϕ) sampled by the GA during the optimization process, where θ and ϕ represent the gantry and couch angles respectively.

However, this inner/outer loop approach is impractical because the total number of evaluations scales as the product of the number of evaluations required for convergence of the inner and outer problems separately. It is more efficient to instead treat beam angle selection and fluence modulation as a single monolithic optimization problem, with all beam angles and fluence parameters handled in exactly the same way, and optimized simultaneously by the GA.

Naturally, a monolithic formulation increases the dimensionality of the optimization problem, which requires an optimizer that performs well on very large problems. Nevertheless, it is beneficial to reduce the number of parameters required to model each fluence pattern where such reduction is possible, since a reduction of dimensionality generally improves mapping of the trade-off surface. At present, we model the fluence pattern as a linear gradient function that is applied to the projection of the PTV in the beam's-eye-view plane:

$$F_{PTV} = F_0 + \left[\mathbf{g} \cdot \left(\mathbf{x} - \mathbf{x}_0\right)\right], \tag{2.1}$$

where $\mathbf{x} = (x, y)$ is a point within the projected PTV, $\boldsymbol{x_0}$ is the isocentre of the projected PTV, $\mathbf{g} = (g_x, g_y)$ is the fluence gradient, and F_0 is the fluence offset. Other methods of parameterizing the fluence pattern are presented in Section 3.2.2. The fluence is set to zero for all pixels that fall outside the projection of the PTV plus a small margin whose width is controlled by an optimization parameter (see Section 3.2.2.7). Furthermore, the fluence is constrained to the range [0, 1] by setting F_{PTV} to zero anywhere that $F_{PTV} < 0$ and to one wherever $F_{PTV} > 1$. A fluence reduction parameter q is specified for each OAR such that

$$F_{OAR} = qF_{PTV}; q \in [0,1] \tag{2.2}$$

for points that fall within the projected boundary of the OAR (see Section 3.2.2.6). Thus, the overall fluence pattern takes the form of a linear truncated gradient over the PTV plus margin, which is reduced by factor q over each OAR, and is zero outside the projected PTV plus margin. An advantage of this approach is that it automatically results in smooth fluence gradients within ROIs, and sharp (discontinuous) gradients at their boundaries. This simple parameterization is not as flexible as the non-parametric approach used by commercial IMRT optimizers to model fluence patterns, and at this early stage of development PARETO's fluence patterns are not expected to be as refined. However, for simple test phantoms, our approach provides sufficient modulation comparable to a commercial treatment planning system as shown by the solutions discussed in Section 2.3.1. Our method may be readily extended to include multiple additive gradients for greater modulation. We are currently investigating this and other more elaborate parameterizations to refine our fluence patterns.

For each beam angle, three parameters are required to model the gradient function, and one additional parameter is required for each OAR. Thus, $(3+N_{OAR})*N_{beam}$ parameters are required to model the fluence pattern, where N_{OAR} is the number of OARs and N_{beam} is the number of beams. The total number of parameters, including beam angles, is given by $N_{par}=(4+N_{OAR})*N_{beam}$ for coplanar beam angle optimizations, and $N_{par}=(5+N_{OAR})*N_{beam}$ for non-coplanar runs. Thirty-five parameters are therefore required for a five beam coplanar problem with three OARs, such as the solutions shown in Section 2.3.2.

Ferret allows parameters to be designated as periodic, such that the minimum and maximum values represent the same configuration. PARETO's gantry angles are treated as periodic, with zero degrees and 360 degrees representing the same position. Couch angles vary from -90 degrees to 90 degrees and are not considered periodic. These ranges map the full sphere containing all possible beam angle orientations. Since the order in which beams are applied does not matter, it is beneficial to reduce the size of the search space by sorting beams for each solution first by gantry angle, and second by couch angle when a pair of beams shares the same gantry angle. This sorting operation is non-trivial because fluence parameters must also be re-arranged within a parameter set when the beam order changes. Ferret was customized for PARETO by adding a re-insertion mechanism that allows these modified parameters to be placed back into the GA population.

PARETO can be run in a mode that allows beams to merge when their angular separation on a sphere defined by (θ, ϕ) is less than a threshold value, which is typically 10 to 15 degrees. PARETO performs a binary search for each solution that identifies and merges beams closer than the threshold angle, in order of increasing angular separation, and stops when no two beams are closer than the threshold. A beam merger averages the vector position of two beams on the unit sphere and also averages their fluence maps prior to the evaluation of the fitness function. Merging results in two identical beams within a solution. However, PARETO ignores duplicate beams by evaluating only one copy of the merged pair within the fitness functions (to see how this affects the relationship between run time and the maximum number of beams, see Appendix B, Fig.

B.1b). It is a subtle point that beam angles modified by merging are evaluated ‘on the fly’ by the fitness function, but are not re-inserted back into the GA population via the parameter re-insertion mechanism used for beam sorting. Merging beams results in the duplication of parameters, which would severely degrade the diversity of parameter values present in the population and damage the search if re-insertion were allowed. Note that beam merging is fundamentally different from beam sorting, which does not result in parameter duplication and helps the search by decreasing the size of the parameter space. The beam merging feature was added to allow preliminary evaluation of PARETO’s ability to optimize the number of beams in addition to beam angles and fluences.

It is a significant challenge to develop realistic multiobjective fitness functions for the OARs and the PTV. These fitness functions are used by the GA as the sole measures of solution quality, and therefore must robustly encode the attributes of the dose distribution as would be judged by a dosimetrist. This assessment must be compressed into a single fitness value for each OAR or PTV. In general, the choice of the OAR and PTV fitness functions affects the character and quality of solutions. Table 1 lists the fitness functions for the OARs and PTV, which are currently implemented in PARETO. We note that Ferret is a minimizer and therefore good solutions correspond to low values of the fitness function. We have studied the behaviour of our fitness functions using the test phantoms described in Section 2.2.3. In the remainder of this section we describe our fitness functions and discuss some of the rationale behind their selection.

Name	Formula	Description
F-OAR-SIMPLE	$F_{OAR,VOL} = \dfrac{N^{-1}\sum_{i=1}^{N} D_{OAR,i}}{M^{-1}\sum_{i=1}^{M} D_{PTV,i}}$ $F_{OAR,SER} = \dfrac{\max\left(\{D_{OAR,i}\}, i \in \{1 \ldots N\}\right)}{M^{-1}\sum_{i=1}^{M} D_{PTV,i}}$	Mean and maximum OAR dose $D_{OAR,i}$, normalized to mean PTV dose $D_{PTV,i}$. N and M are the number of OAR and PTV voxels respectively. $F_{OAR,VOL}$ is used for volumetric structures, and $F_{OAR,SER}$ is used for serial structures.
F-OAR-DVH	$F_{OAR} = \dfrac{1}{V_{OAR}}\left[\sum_{i=1}^{N} f_i^p \Delta V_{OAR,i}\right]^{1/p}$ $f_i = \begin{cases} 0, & D_i \le D_0 \\ D_i / D_0, & D_i > D_0 \end{cases}$	Volume-weighted mean thresholded dose computed from a dose-volume histogram (DVH) with N bins. $\Delta V_{OAR,i}$ is the increment in volume from the i'th bin of the DVH curve and V_{OAR} is the total volume. D_i is dose and D_0 is the target dose threshold for the OAR. Voxels receiving the highest dose are more strongly suppressed for larger values of exponent $p \ge 1$.
F-OAR-EUD	$F_{EUD} = \left\langle \sum_{i=1}^{N} f_i^p \right\rangle^{1/p}$ $f_i = \begin{cases} 0, & D_i \le D_0 \\ D_i / D_0, & D_i > D_0 \end{cases}$	Equivalent uniform dose with dose threshold D_0. N is the number of OAR voxels. Exponent $p = 1$ best corresponds to volumetric structures. Values of $p > 1$ are most appropriate for serial structures. Other quantities are as defined in F-OAR-DVH.
F-PTV-STD	$F_{PTV} = \dfrac{\left[\sum_{i=1}^{M}\left(D_i - \bar{D}\right)^2\right]^{1/2}}{\sum_{i=1}^{M} D_i}$	Normalized standard deviation of dose over PTV region. D_i is dose within the PTV and $\bar{D}$ is mean dose.
F-PTV-CONF and F-PTV-CONF-HOTSPOT	$F_{PTV} = \log_{10}\left\{\dfrac{1}{2}\left[1 - \dfrac{\sum_{i=1}^{N} Q_i\left(D_i - \bar{D}\right)\left(M_i - \bar{M}\right)}{\sigma_D \sigma_M}\right]\right\}$ $\sigma_D = \left(\dfrac{\sum_{i=1}^{N} Q_i\left(D_i - \bar{D}\right)^2}{\sum_{i=1}^{N} Q_i}\right)^{1/2}$ $\sigma_M = \left(\dfrac{\sum_{i=1}^{N} Q_i\left(M_i - \bar{M}\right)^2}{\sum_{i=1}^{N} Q_i}\right)^{1/2}$	Conformity fitness function based on weighted statistical correlation function between dose D and a PTV mask M, defined such that a value of 1 is assigned to all voxels inside the PTV and 0 outside. σ_D and σ_M are the weighted dose and mask standard deviations, respectively. N is the total number of voxels in the mask. Weight matrix Q_i determines the relative weight applied interior and exterior to the PTV. The F-PTV-CONF-HOTSPOT fitness function is identical to F-PTV-CONF, except that the mask M is restricted to a "hot zone" near the PTV, as discussed in Section 2.2.2.

Table 1: The fitness functions implemented within PARETO. A setup file allows the user to select OAR and PTV fitness functions. A minimization problem is assumed, so that smaller values of the fitness function correspond to superior solutions for the corresponding objective.

Earlier versions of PARETO used a combination of the PTV dose standard deviation (F-PTV-STD) and simple OAR mean dose (F-OAR-SIMPLE) fitness functions. The F-OAR-SIMPLE function places too much emphasis on reducing dose below normal thresholds, and thus very few balanced solutions were discovered using this function. Solution sets typically contained a few good solutions, but these were diluted by a larger number of poor solutions characterized by extremely low dose for one or more OARs, but unacceptably high dose for other OARs and poor PTV uniformity. This problem is remedied by fitness functions that include a dose threshold D_0 for each OAR, as discussed below. The F-PTV-STD fitness function resulted in a poor spatial distribution of beams. With this PTV fitness function, PARETO typically allowed too many beams to overlap or merge, resulting in defective solutions characterized by poor conformity to the PTV. Severe hotspots of dose exterior to the PTV were not suppressed because the uniformity fitness function computes the standard deviation of the dose over the PTV only, without considering voxels outside.

The PTV conformity fitness function (F-PTV-CONF) is based on the statistical correlation function between the dose D and a mask M, over all voxels within the treatment region. The mask corresponds to a perfect dose distribution, with a value of one assigned to voxels within the PTV, and a value of zero assigned outside the PTV. This function attempts to achieve a uniform dose distribution within the PTV structure and minimal uniform dose outside, with a steep gradient at the PTV boundary. The conformity function is designed to allow variable weighting between the PTV and the exterior. Weighting was included because we find in general that equal weighting between the PTV and exterior overemphasizes the exterior, resulting in degraded PTV uniformity. This problem is alleviated by placing more weight on the PTV than the exterior region by setting:

$$w_i = \begin{cases} w_{PTV}, & \text{PTV voxels} \\ 1 - w_{PTV}, & \text{outside} \end{cases} , \tag{2.3}$$

where we have chosen $w_{PTV} = 5/7$ in this work. It is possible within our framework to eliminate this weighting by splitting the conformity fitness function into two separate uniformity objectives for regions interior and exterior to the PTV. However, this increase in the number of objectives is not warranted because our solutions are only weakly sensitive to w_{PTV}.

Better spacing of beams was noted for the F-PTV-CONF and F-OAR-SIMPLE combination of fitness functions compared to those found using the F-PTV-STD and F-OAR-SIMPLE combination. However, very few balanced solutions were present in the non-dominated set. Some OARs were driven to very low dose at the expense of others, and the PTV uniformity was poor for most solutions.

Dose-volume histograms (DVHs) are used universally to assess the quality of treatment plans. Therefore, PARETO implements a DVH-based fitness function (F-OAR-DVH in Table 1) to simulate this decision process. This fitness function is based on the mean DVH-weighted dose, normalized to the mean PTV dose and modified by dose threshold D_0 for each OAR. Results are dependent on the dose threshold, which must be chosen carefully to reflect realistic targets for OARs. Choosing thresholds that are too low results in a non-dominated set populated with many clinically unacceptable solutions with dose targets that cannot be achieved and poor PTV uniformity, much like our experiments with the F-OAR-SIMPLE function discussed above. Naturally, we would prefer more general fitness functions that do not require dose thresholds as extra free (non-optimized) parameters. However, well-chosen thresholds are necessary in the present implementation. Additionally, an exponent p affects the emphasis placed on suppressing high dose voxels, such that $p = 1$ is an appropriate choice for organs with parallel functional subunit (FSU) structure, while $p > 1$ more strongly suppresses small regions of high dose, and therefore may be more appropriate for organs with serial FSU structure (see Section 1.2).

PARETO also implements a fitness function labeled as F-OAR-EUD in Table 1, which is based on the equivalent uniform dose [33]. The similarity to the simple OAR fitness function is obvious, except that a dose threshold D_0 and a dose exponent p are implemented in the same manner as the DVH fitness function. The normalized equivalent uniform dose function is more biologically motivated than the DVH fitness function, although the DVH function might more closely reflect the decision process of a clinician. Both produce solutions of excellent quality that are of similar character as discussed in Section 2.3.2.

In addition to the fitness functions discussed above, we include an optional penalty function designed to explicitly suppress hotspots distant from the PTV, which is currently added to all other fitness functions. We note, however, that our study of the relatively poorly performing F-PTV-STD and F-OAR-SIMPLE functions preceded this modification, and have not been tested with the hotspot penalty function. The penalty function is defined by the equations:

$$
\begin{aligned}
&dose\_violation = \max\left\{0, \left(\frac{D_i - D_{OHT}}{D_{OHT}}\right)\right\} \quad \forall i \in \{1 \ldots N\} \middle| d_i > f_{PTV} * d_{PTV} \\
&F_{penalty} = \begin{cases} 0, & if\ dose\_violation = 0 \\ F_0 + dose\_violation, & if\ dose\_violation > 0 \end{cases}
\end{aligned}
\tag{2.4}
$$

where d_{i} is the distance from voxel i within the treatment volume to the nearest voxel within the PTV structure, and d_{PTV} represents the maximum distance between points within the PTV. The product f_{PTV} * d_{PTV} defines the width of a "hot zone" near the PTV, where f_{PTV} is a fractional value typically set to approximately 0.5. The typical scale of the penalty is given by F_0, which is chosen to be greater than typical values of all other fitness functions. The radiation dose is allowed to remain high within the hot zone, which is necessary due to the intersection of beams in proximity to the PTV. However, the penalty is applied if there exists a voxel outside the hot zone that receives a normalized dose D_i greater than the allowable other healthy tissue (OHT) dose D_{OHT}. The penalty increases linearly with the fractional degree by which the allowable OHT dose is exceeded. Note that the penalty function is equal to zero if there are no hot voxels outside the hot zone. The early part of a PARETO run quickly learns to avoid regions of parameter space that results in dose hotspots, but the fitness functions are unaffected once all hotspots have been eliminated. Our tests show that this penalty function is very effective at excluding solutions with problematic hotspots.

When the hotspot penalty function is used, a small additional improvement is possible by restricting the mask *M* used in the F-PTV-CONF fitness function to include only voxels within the PTV and hotspot region. The rationale for this modification is that the fitness function can then focus on optimizing the dose distribution near to the PTV, while the hotspot penalty function eliminates solutions containing problematic hotspots distant from the PTV. This modified function is referred to as F-PTV-CONF-HOTSPOT in Table 1. We have used this fitness function in combination with the hotspot penalty for all runs shown in Section 2.3, using $D_{OHT} = 0.8$, $f_{PTV} = 0.5$, and $F_0 = 10$.

The preferred OAR fitness functions within PARETO (F-OAR-DVH and F-OAR-EUD) both implement a target dose threshold, which appears to be the main factor responsible for their superiority compared to the F-OAR-SIMPLE function. The GA works to decrease OAR fitness values early in the run, and then focuses mainly on the PTV fitness function late in the run. This results in a large population of well-balanced solutions in the final non-dominated set that simultaneously achieve realistic dose thresholds and maximize PTV uniformity.

2.2.3 Description of phantoms used for testing

Several test phantoms have been used during PARETO's development, two of which will be examined here. The first phantom is a simple homogeneous cylindrical phantom (Fig. 2.3) with a central cylindrical target surrounded by three cylindrical OARs. The phantom is 20 cm in diameter and 10 cm in height. The PTV has a diameter of 5.1 cm and a height of 6 cm, while the OARs all have diameters of 2 cm and heights of 6 cm. The second phantom is more complex and closer represents a realistic clinical situation. As illustrated in Figure 2.4, this phantom was designed to represent a paraspinal patient, with OARs consisting of both left and right kidneys as well as the spinal cord, with the target defined as a C-shaped structure wrapping around the spinal cord.

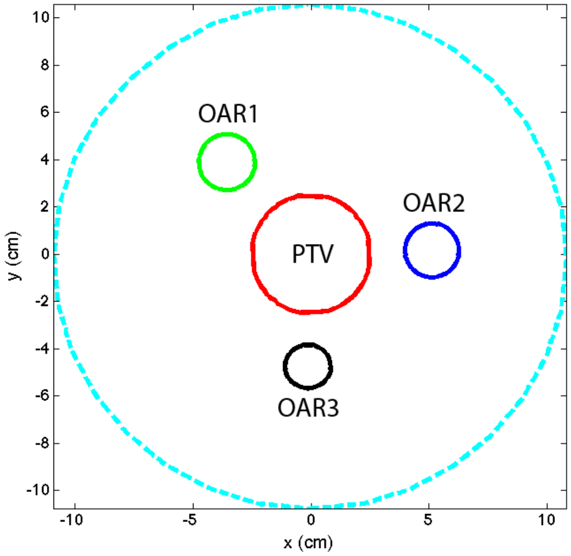

Figure 2.3: An axial slice of a homogeneous cylindrical phantom. The phantom is defined on 13 slices in the z direction with the PTV and OARs occupying the central 7 slices. The colours correspond to the DVH curves shown in Figures 2.5-2.7.

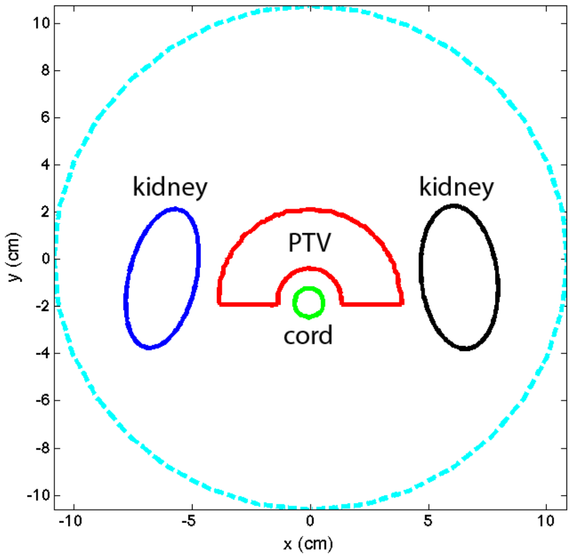

Figure 2.4: An axial slice of a spinal phantom comprised of a C-shaped PTV wrapped around an OAR representing a spinal cord, between two other OARs represent kidneys. The phantom is defined on 13 slices in the z direction with the PTV and OARs occupying the central 7 slices. The colours correspond to the DVH curves shown in Figure 2.10.

2.2.4 Methodology for Comparison with a Commercial Treatment Planning System

This section describes the methodology used to compare solutions from PARETO with the Eclipse treatment planning system (Varian Medical Systems, Palo Alto, CA). PARETO's multiobjective optimization approach is fundamentally different from the single-objective approach used in existing commercial treatment planning systems such as Eclipse. In addition, the fitness functions in Table 1 are designed specifically for PARETO, and differ from the weighted terms used for single-objective optimization. We therefore expect a correlation between DVHs found by PARETO and commercial systems, but not exact quantitative agreement, since different objectives have been optimized. In order to compare PARETO solutions with those computed from Eclipse, we search through the PARETO solution database to find a multiobjective solution that achieves the set dose threshold D_0 for all OARs, and then we select the one with the lowest PTV fitness evaluation. We then define the objective dose as D_0 for each OAR in Eclipse, input the beam angles from the selected PARETO solution, and optimize the fluence. We export the DVH data from Eclipse for comparison with DVH curves computed using PARETO.

Several metrics are used to evaluate plan quality including: the standard deviation of dose within the PTV structure, the maximum point dose (hot spot) outside the PTV structure, and the radiation conformity index [34] of the 95% isodose surface, RCI_{95} (defined as the volume of the PTV divided by the volume of the 95% isodose surface, with the 95% isodose surface completely encompassing the PTV).

2.3 Results

For the results presented here, we used an exponent of $p = 1$ for the F-OAR-DVH and F-OAR-EUD fitness functions, in combination with the F-PTV-CONF-HOTSPOT fitness function and the hotspot penalty function. A total population size of 500 was used for the GA, which optimized for 1000 generations, requiring approximately 15 hours of computation in parallel on a server with twelve CPU cores. We note that such long run times would not be necessary in clinical practice, since runs converge after the first 100 to 200 generations, with no further improvement in OAR fitness values and only very small improvements of a few percent in the PTV fitness. The non-dominated set was essentially stable at this point, with the remaining generations serving to increase the number of solution points and provide slight refinement. For testing purposes, we often run PARETO for a few hundred generations using population sizes of only 100 to 200, which decreases the run time by approximately an order of magnitude. The resulting trade-off surfaces are sparser than the ones shown here, but contain plans of similar quality.

2.3.1 Comparison with a Commercial Treatment Planning System

Figure 2.5 demonstrates a good correlation between PARETO and Eclipse dose-volume histograms for runs performed using dose thresholds of D_0 = 0.15, 0.20, and 0.30. For each threshold, we selected an optimized solution from the trade-off surface, and the PARETO-optimized beam angles were input manually into Eclipse for comparison. PARETO appears to find solutions offering comparable sparing of the OARs and superior uniform coverage of the PTV. However, we note that PARETO does not yet take into account multi-leaf collimator (MLC) sequencing, which would cause some degradation of solution quality. The comparison between PARETO and Eclipse may be strengthened by improving PARETO's dose calculation accuracy through the use of a two-source linear accelerator head fluence model including MLC modeling, and conversion of ideal fluence maps to deliverable MLC sequences.

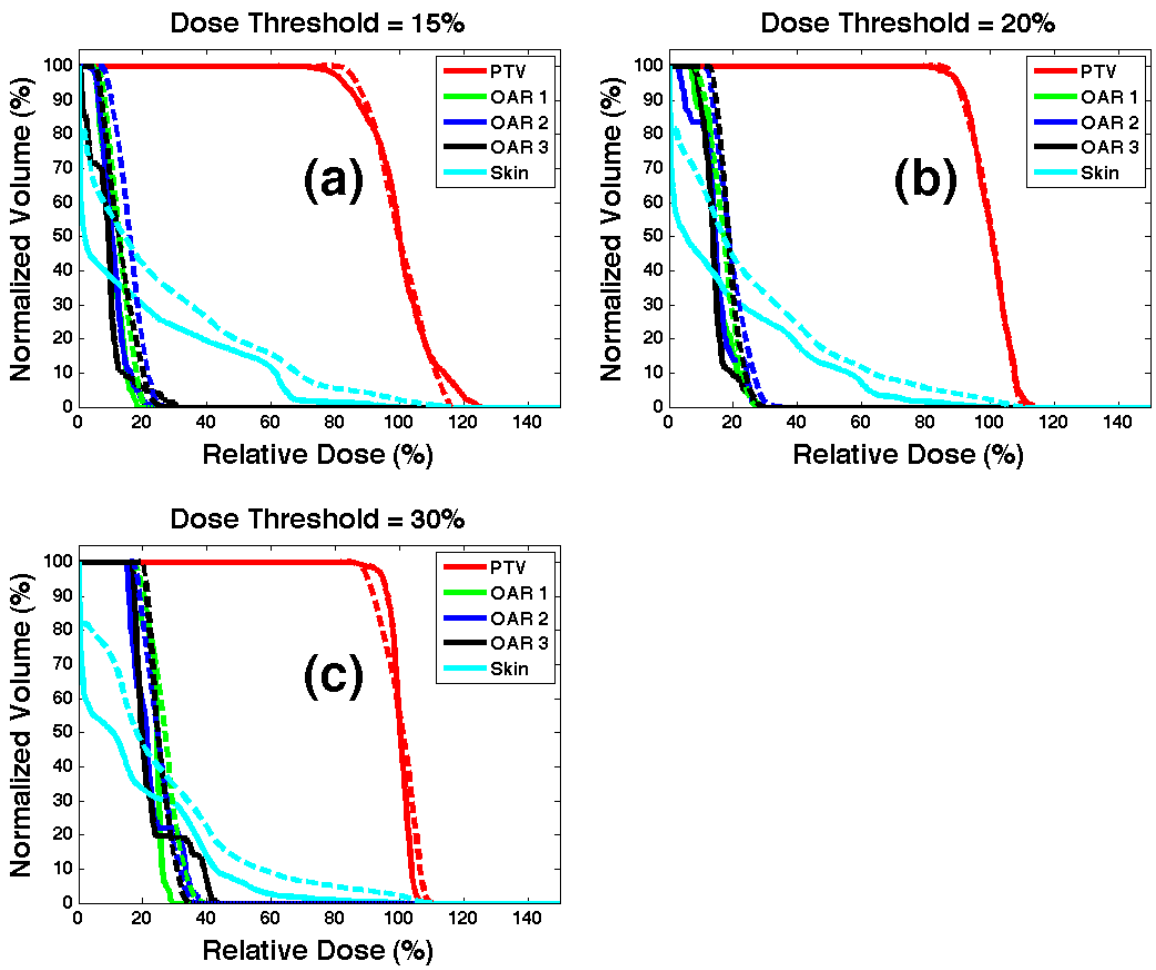

Figure 2.5: DVH curves for PARETO (solid line) and Eclipse (dashed line) solutions computed for the cylindrical phantom using dose thresholds of (a) 15% (b) 20% and (c) 30% of the prescription dose. The colours correspond to the structures shown in Figure 2.3.

2.3.2 Comparison of Fitness Functions

2.3.2.1 OAR DVH Fitness Function with PTV Conformity Fitness

Figure 2.6 shows the final non-dominated set projected into the plane of the F-PTV-CONF-HOTSPOT fitness function and the quadrature-averaged F-OAR-DVH fitness functions. We note that the quadrature average is used for visualization purposes only; the OAR fitness functions were not combined during the multiobjective optimization. Each data point represents a non-dominated plan. Dose distributions and DVHs for three highlighted plans (A, B, C) are shown. A dose threshold of $D_0 = 0.20$ was used. The maximum OHT point dose relative to the PTV prescribed dose was 102%, 112%, 105%, the normalized PTV dose standard deviation was 10.9%, 6.94%, 3.97%, and the radiation conformity index of the 95% isodose volume (RCI_{95}) was 1.27, 1.21, 1.05 for solutions A, B, C, respectively. Using this combination of fitness functions, the PTV dose is relatively uniform and conformal, beams are well spread out, and no hotspots are apparent (Fig. 2.6).

2.3.2.2 Normalized OAR EUD Fitness Function with PTV Conformity Fitness

We find that solutions using the EUD fitness function (Fig. 2.7) are very similar to those computed using the DVH fitness function. The maximum OHT dose was 103%, 105%, 97.0%, the PTV dose standard deviation was 7.89%, 4.68%, 4.97%, and the RCI_{95} was 1.22, 1.06, 1.09, for solutions A, B, C, respectively. In this case, a dose threshold of $D_0 = 0.20$ was used, however, we note that the EUD fitness function reduces to the simple OAR fitness function when $p = 1$ and $D_0 = 0$, which confirms that the OAR dose threshold is the main factor responsible for the greatly improved solution quality over the simple OAR fitness function.

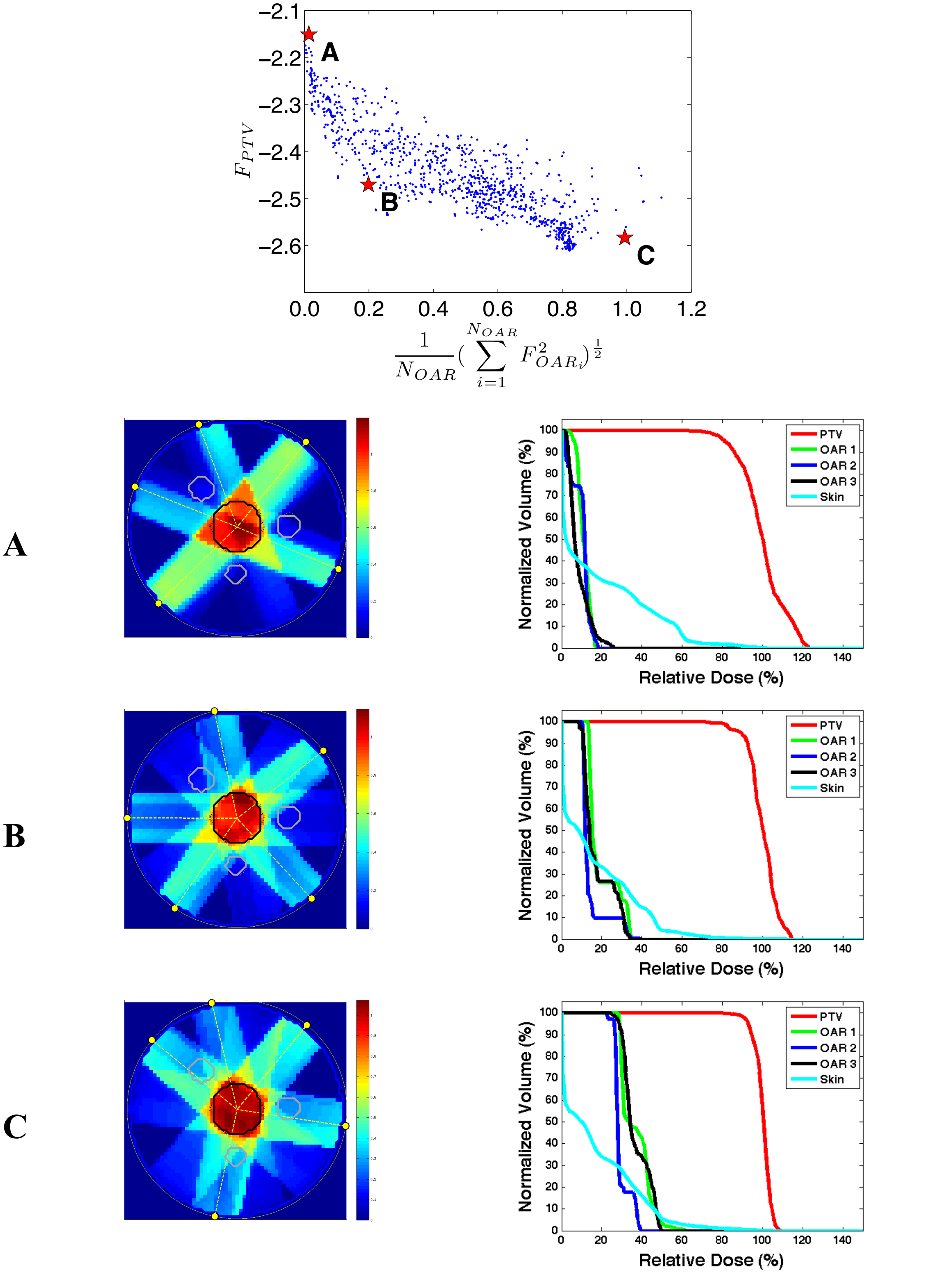

Figure 2.6: The Pareto non-dominated set for the cylindrical phantom as projected on the plane of the F-PTV-CONF-HOTSPOT fitness function and the F-OAR-DVH (dose threshold = 20%) fitness functions added in quadrature. Dose distributions and DVHs for three highlighted plans (A, B, C) are shown. The PTV and OARs are contoured, and colours on the DVH curves correspond to the structures shown in Figure 2.3. The dashed lines indicate beam angles.

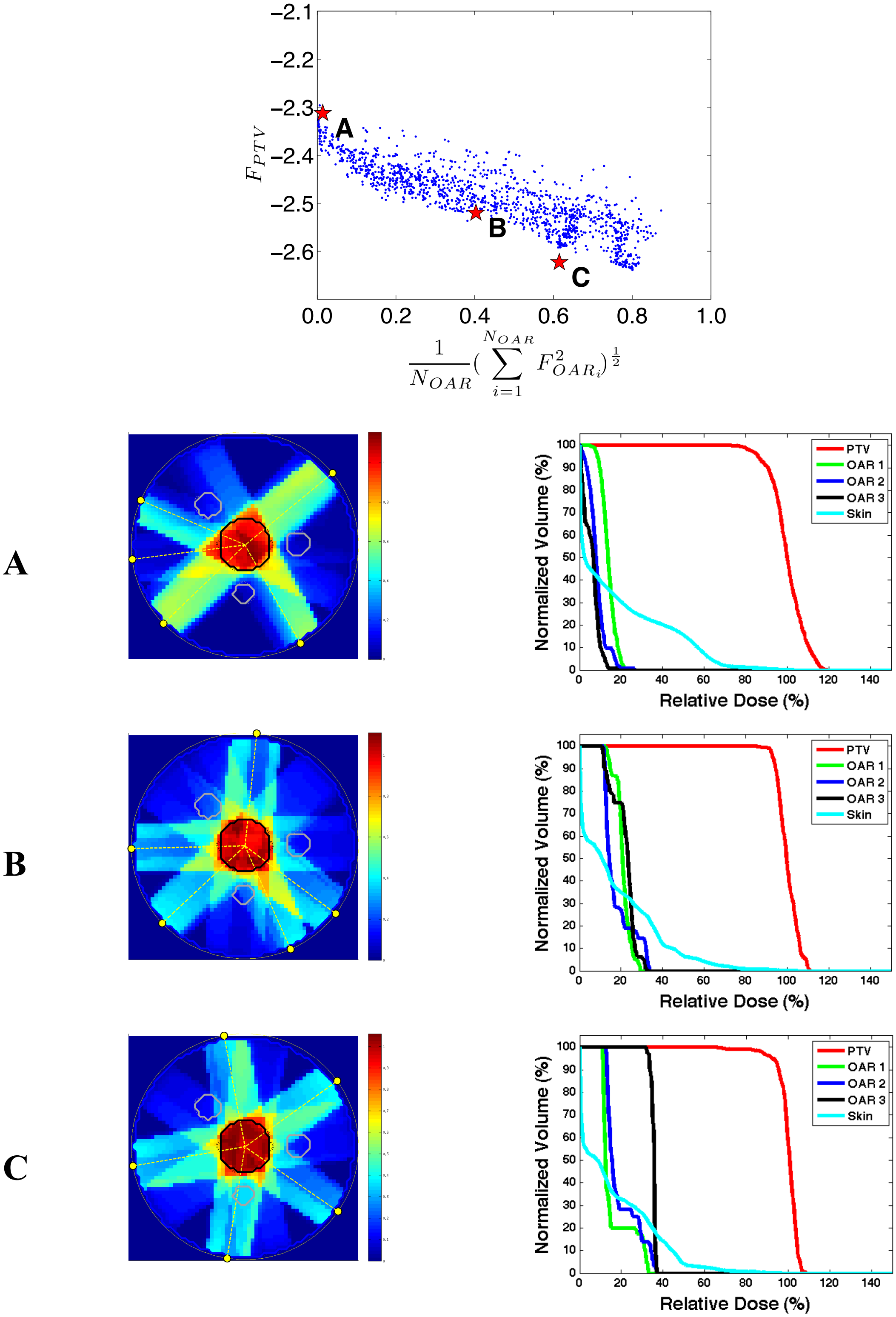

Figure 2.7: The Pareto non-dominated set for the cylindrical phantom as projected on the plane of the F-PTV-CONF-HOTSPOT fitness function and the F-OAR-EUD (dose threshold = 20%) fitness functions added in quadrature. Dose distributions and DVHs for three highlighted plans (A, B, C) are shown. The PTV and OARs are contoured, and colours on the DVH curves correspond to the structures shown in Figure 2.3. The dashed lines indicate beam angles.

2.3.2.3 Beam Number Optimization

In this Section, we show results from a run using the cylindrical phantom and a maximum of 11 beams, with beam merging enabled such that beams closer than 15 degrees were combined. PARETO generated a database of 2325 non-dominated solutions using the F-OAR-DVH and F-PTV-CONF-HOTSPOT fitness functions with a dose threshold of $D_0 = 0.20$. A slight perturbation $\Delta F = 10^{-6} N_{beams}$ was added to all fitness functions to slightly prefer solutions with fewer beams, where N_{beams} is the number of unique beams after merging. It is important to note that the Ferret GA uses tournament Pareto ranking; thus, this tiny perturbation is sufficient to strongly prefer solutions with fewer beams late in the run when the population is rich with solutions that achieve their dose thresholds such that $F_{OAR} = 0$ for one or more OARs.

Figure 2.8a shows the final non-dominated set projected into the plane of the F-PTV-CONF-HOTSPOT fitness function and the quadrature-averaged F-OAR-DVH fitness functions. Four plans containing eleven, nine, seven, and five unique beam angles are highlighted. Figure 2.8b demonstrates a trend of improved dose conformity to the PTV by increasing the number of optimized beams. The maximum OHT dose was 101%, 110%, 104%, 108%, the PTV dose standard deviation was 3.99%, 5.08%, 6.01%, 7.99%, and the RCI_{95} was 1.06, 1.10, 1.21, 1.17, for the eleven, nine, seven and five beam plans, respectively. We note that the PTV fitness function is logarithmic, and as such the actual differences in the conformity function are small for highly conformed solutions toward the maximum number of beams shown. Figure 2.8b therefore suggests that PTV conformity gradually saturates for a high number of beams, which is a result similar to earlier work examining the optimal number of beam angles for IMRT [35].

Figure 2.9 compares DVH curves for solutions containing different numbers of beams. Panel a) shows the trend of improving PTV uniformity with increasing number of beams. However, no obvious trend is noted for OARs.

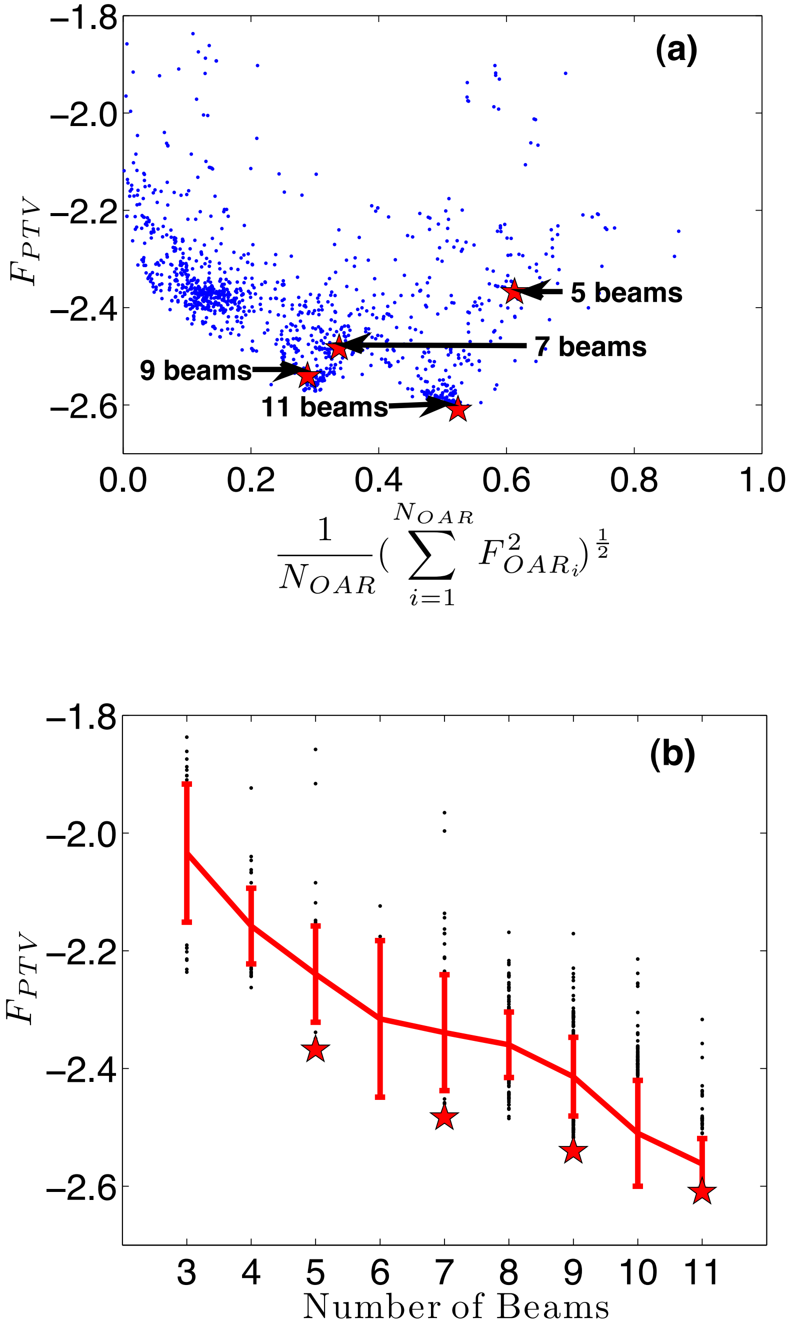

Figure 2.8: (a) The Pareto non-dominated set for an 11-beam run on the cylindrical phantom, as projected on the plane of the F-PTV-CONF-HOTSPOT fitness function and the F-OAR-DVH fitness functions averaged in quadrature. (b) The F-PTV-CONF-HOTSPOT fitness function values for plans with a varying number of beams. Red lines join the mean fitness value of each bin. The standard deviation of the fitness values within each bin is shown as an error bar. Four plans containing eleven, nine, seven and five beams are highlighted.

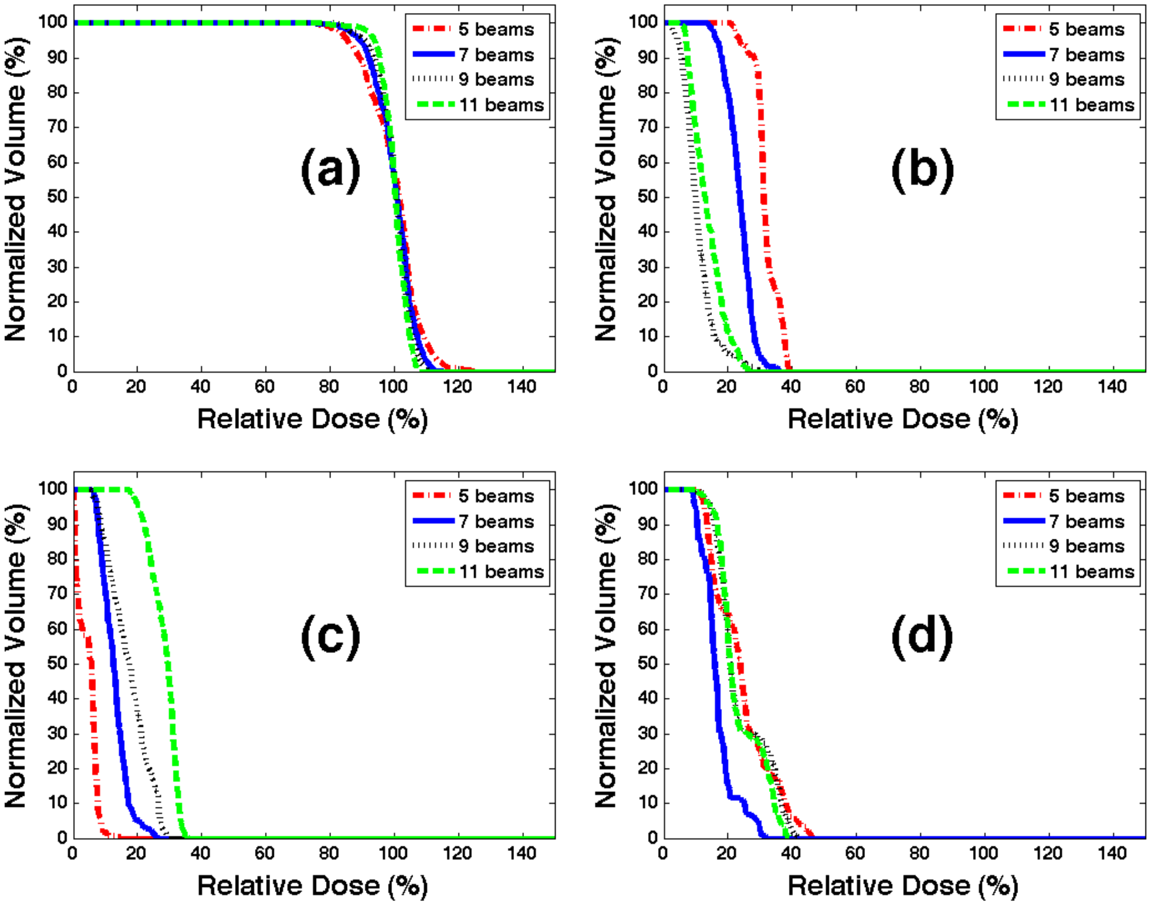

Figure 2.9: Dose-volume histograms for an 11-beam run on the cylindrical phantom showing (a) the PTV, (b) OAR 1, (c) OAR 2, and (d) OAR 3 for the highlighted plans in Figure 2.8.

2.3.2.4 Results for a Spinal Phantom

PARETO was run for 1000 generations on the spinal phantom shown in Figure 2.4 using a dose threshold of D_0 = 0.20, resulting in 2633 non-dominated solutions. We specified a maximum of 7 beams and allowed them to merge according to the beam number optimization method described in Section 2.3.2.3. Figure 2.10 shows the projection of the F-PTV-CONF-HOTSPOT fitness function and the quadrature-averaged F-OAR-DVH fitness functions and three plans selected from the trade-off surface. The maximum OHT dose for this run was 127%, 122%, 107%, the PTV dose standard deviation was 12.1%, 7.45%, 5.27%, and the RCI_{95} was 1.08, 1.09, 1.08, for solutions A, B, and C respectively.

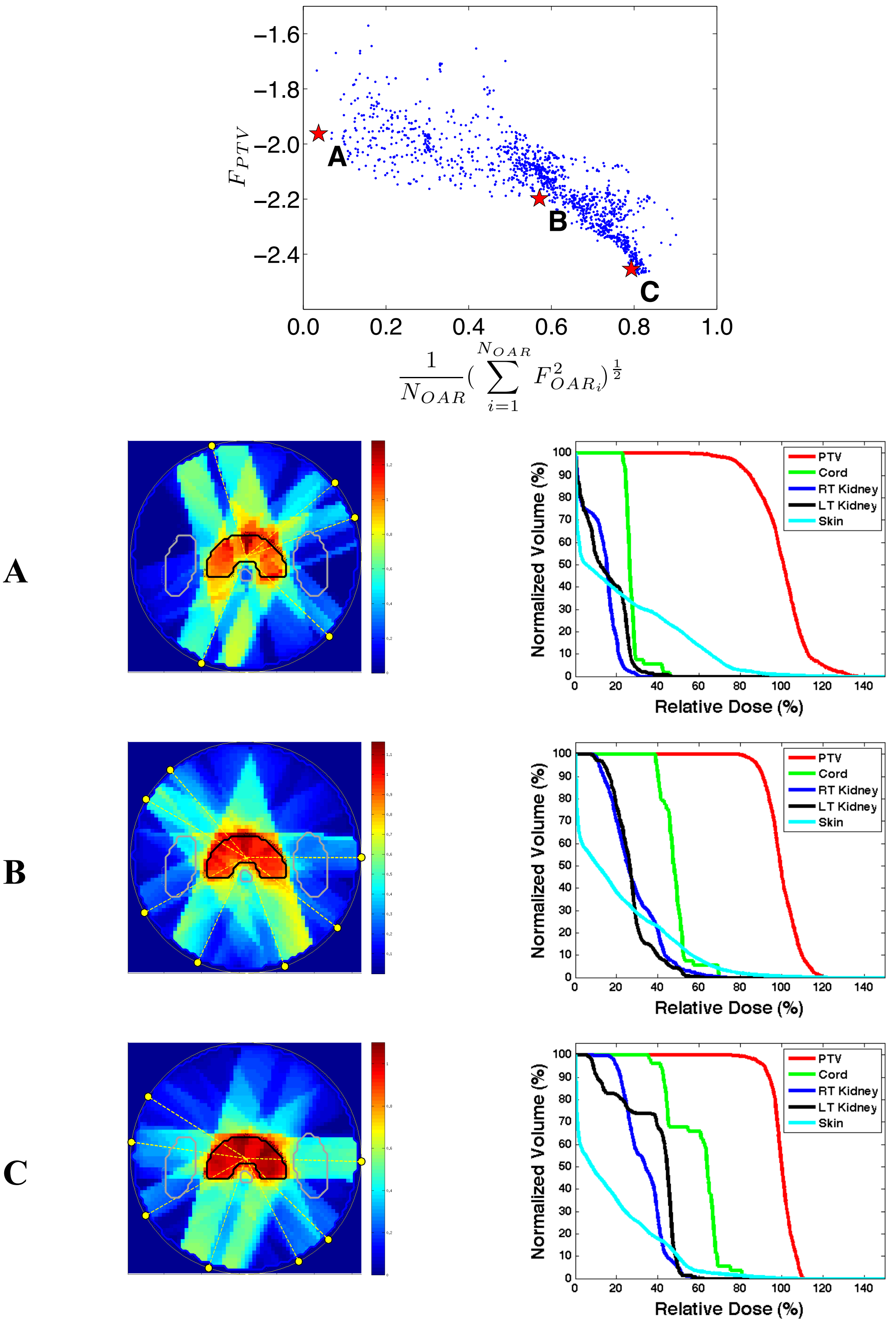

Figure 2.10: The Pareto non-dominated set for the spinal phantom as projected into the plane of the F-PTV-CONF-HOTSPOT fitness function and the F-OAR-DVH (dose threshold = 30%) fitness functions added in quadrature. The hotspot suppression penalty function was used. Dose distributions and DVHs for three highlighted plans (A, B, C) are also shown. The PTV and OARs are contoured, and colours on the DVH curves correspond to the structures shown in Figure 2.4. The dashed lines indicate beam angles.

2.4 Discussion

This work demonstrates that PARETO is able to find high quality solutions to the problem of multiobjective IMRT treatment planning. PARETO does not yet optimize fluence patterns as finely as commercial treatment planning systems, but the software's rudimentary linear gradient method is sufficiently flexible to permit the discovery of good beam orientations. Moreover, solutions are rapid to evaluate, and therefore suitable for global optimization with a GA. More sophisticated fluence parameterizations are under development to further improve the PTV dose uniformity and OAR sparing. Our fluence parameterization resulted in DVH curves that are comparable to those calculated by Eclipse. However, this comparison must be approached with caution until PARETO is able to model the linear accelerator head fluence, including the MLC, and conversion of optimal fluences to deliverable MLC sequences. Future work will explore these important details.

The end result of a PARETO run is a database of non-dominated solutions that shows trade-offs between OAR and PTV fitness functions, typically containing at least 10^3 solutions, which are all equally good in the Pareto-optimal sense. Thus, good interactive visualization tools are essential for the physician to navigate the database to select the best plan to treat the patient. We have implemented a preliminary "PARETO Navigator" tool for this purpose, which will be discussed elsewhere. Given the fact that plan optimization does not require human supervision, and the reasonable assumption that the selection of the best plan from the trade-off surface can be made rapidly through a well designed graphical user interface, this implies that PARETO decreases the time required of a human treatment planner, and therefore may increase patient throughput.

The multiobjective tool developed here has two major implications for clinical treatment planning. The first is that simultaneously optimizing beam angles and fluence patterns by solving the full multiobjective problem may result in better quality solutions than those developed by manual selection of beam angles followed by fluence optimization. The second is that it is possible to streamline treatment planning by providing an automated multiobjective technique, which eliminates the need for iterative, human manipulation of weighting factors and dose objectives during optimization. We emphasize that PARETO does not eliminate the need for human clinical input during treatment planning, but rather shifts the focus to rapid exploration of a pre-optimized set of solutions and expert selection of a treatment plan.

PARETO makes use of Ferret's internal parallelization and run times scale approximately as $N_{CPU}^{-0.8}$, where N_{CPU} is the number of CPU cores employed [6]. Long run times of approximately 15 hours were used to compute the densely sampled trade-off surfaces shown in this paper. However, less densely populated trade-off surfaces containing plans of similar quality can be produced from much shorter runs. Clinically viable optimization times are within reach, given a sufficient number of cores, noting that the optimization can be performed offline, leaving the human operator free for other tasks. In addition, we are testing a new version of PARETO that uses several graphical processing units (GPUs) to increase the speed of computation. The GPU work will be discussed in a future publication and benchmarked against the current version, which used CPU-based computation only.

2.5 Conclusions

We have introduced PARETO, a multiobjective optimization tool to solve the IMRT treatment planning problem. PARETO utilizes a powerful evolutionary algorithm to handle the combined monolithic problem of beam fluence and beam angle optimization. We have employed a simple parameterized beam fluence representation and a realistic dose calculation to demonstrate feasibility for a simple cylindrical phantom and a more realistic paraspinal phantom. The combination of the conformity-based PTV fitness function (F-PTV-CONF-HOTSPOT), the DVH or EUD-based fitness functions for the OARs, and the hotspot penalty function produced acceptably uniform and conformal PTV doses, with well-spaced beams and no hotspots. Results also indicated that PARETO shows promise in optimizing the number of beams. A DVH comparison of solutions from PARETO to those developed with a single-objective fluence optimizer (commercial planning system Eclipse) showed a good correlation. Work is underway to develop improved fluence parameterizations that are more flexible than the simple gradient-based method discussed here. Future versions of PARETO will include a linear accelerator head fluence model and MLC sequencer, as the software evolves toward clinical use.

2.6 Acknowledgements

The authors gratefully acknowledge support from the Winnipeg Health Sciences Centre Foundation, the Manitoba Health Research Council, the Natural Sciences and Engineering Research Council of Canada, and CancerCare Manitoba. The authors disclose that they have filed a patent application for PARETO. J. Fiege also discloses authorship of the Qubist Global Optimization Toolbox and ownership of nQube Technical Computing Corporation, which distributes this optimization software.

2.7 References

[1] Memorial Sloan Kettering Cancer Center, *A Practical Guide to Intensity-Modulated Radiation Therapy.* Madison, WI: Medical Physics Publishing, 2003.

[2] J. VanDyk, *Modern Technology of Radiation Oncology Volume 2*. Madison, WI: Medical Physics Pubishing, 2005.

[3] K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms.* John Wiley & Sons, ISBN - 10:047187339X, ISBN - 13:9780471873396, 2001.

[4] J. O. Deasy, "Multiple local minima in radiotherapy optimization problems with dose-volume constraints," *Med. Phys.*, 24 (7), pp. 1157-61, 1997.

[5] A. B. Pugachev, A.L. Boyer, and L. Xing, "Beam orientation optimization in intensity-modulated radiation treatment planning," *Med. Phys.*, 27 (6), pp. 1238-45, 2000.

[6] J. D. Fiege, *Qubist User's Guide: Optimization, Data-Modeling, and Visualization with the Qubist Global Optimization Toolbox for MATLAB.* Winnipeg: nQube Technical Computing Corp., 2010.

[7] J. F. Aubry, et al., "Multiobjective optimization with a modified simulated annealing algorithm for external beam radiotherapy treatment planning*", Med. Phys.*, 33 (12), pp. 4718-29, 2006.

[8] S. Breedveld, et al., "A novel approach to multi-criteria inverse planning for IMRT," *Phys. Med. Biol.*, 52 (20), pp. 6339-53, 2007.

[9] C. Cotrutz, et al., "A multiobjective gradient-based dose optimization algorithm for external beam conformal radiotherapy," *Phys. Med. Biol.*, 46 (8), pp. 2161-75, 2001.

[10] D. Craft, and T. Bortfeld, "How many plans are needed in an IMRT multiobjective plan database?," *Phys. Med. Biol.*, 53 (11), pp. 2785-96, 2008.

[11] D. Craft, T. Halabi, and T. Bortfeld, "Exploration of tradeoffs in intensity-modulated radiotherapy*,"* *Phys. Med. Biol.*, 50 (24), pp. 5857-68, 2005.

[12] D. Craft, et al., "An approach for practical multiobjective IMRT treatment planning," *Int. J. Radiat. Oncol. Biol. Phys.*, 69 (5), pp. 1600-7, 2007.

[13] D. Craft, and M. Monz, "Simultaneous navigation of multiple Pareto surfaces, with an application to multicriteria IMRT planning with multiple beam angle configurations," *Med. Phys.*, 37 (2), pp. 736-41, 2010.

[14] S. K. Das, "A method to dynamically balance intensity modulated radiotherapy dose between organs-at-risk," *Med. Phys.*, 36 (5), pp. 1744-52, 2009.

[15] J. D. Fenwick and J. Pardo-Montero, "Homogenized blocked arcs for multicriteria optimization of radiotherapy: analytical and numerical solutions," *Med. Phys.*, 37 (5), pp. 2194-206, 2010.

[16] A. L. Hoffmann, et al., "Derivative-free generation and interpolation of convex Pareto optimal IMRT plans," *Phys. Med. Biol.*, 51 (24), pp. 6349-69, 2006.

[17] T. S. Hong, et al., "Multicriteria optimization in intensity-modulated radiation therapy treatment planning for locally advanced cancer of the pancreatic head," *Int. J. Radiat. Oncol. Biol. Phys.*, 72 (4), pp. 1208-14, 2008.

[18] M. Monz, et al., "Pareto navigation: algorithmic foundation of interactive multi-criteria IMRT planning," *Phys. Med. Biol.*, 53 (4), pp. 985-98, 2008.

[19] J. Pardo-Montero and J.D. Fenwick, "An approach to multiobjective optimization of rotational therapy," *Med. Phys.*, 36 (7), pp. 3292-303, 2009.

[20] E. Schreibmann, et al., "Multiobjective evolutionary optimization of the number of beams, their orientations and weights for intensity-modulated radiation therapy," *Phys. Med. Biol.*, 49 (5), pp. 747-70, 2004.

[21] E. Schreibmann and L. Xing, "Feasibility study of beam orientation class-solutions for prostate IMRT," *Med. Phys.*, 31 (10), pp. 2863-70, 2004.

[22] T. Spalke, D. Craft, and T. Bortfeld, "Analyzing the main trade-offs in multiobjective radiation therapy treatment planning databases," *Phys. Med. Biol.*, 54 (12), pp. 3741-54, 2009.

[23] C. Thieke, et al., "A new concept for interactive radiotherapy planning with multicriteria optimization: first clinical evaluation," *Radiother. Oncol.*, 85 (2), pp. 292-8, 2007.

[24] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning.* Toronto: Addison-Wesley, 1989.

[25] D. E. Goldberg, *The Design of Innovation.* Boston: Kluwer Academic Publishers, 2002.

[26] C. M. Fonseca, "Genetic algorithms for multiple objective optimization: formulation, discussion and generalization," in *Proceedings of the Fifth International Conference on Genetic Algorithms [ICGA5],* S. Forrest, Ed. Urbana-Champaign, IL: Morgan Kaufmann, pp. 416-423, 1993.

[27] J. D. Fiege, et al., "A genetic algorithm-based exploration of three filament models: a case for the magnetic support of the G11.11-0.12 infrared-dark cloud," *Ap. J.*, 616, pp. 925-942, 2004.

[28] J. D. Fiege, "Computational intelligence techniques for submillimetre polarization modeling," in *Astronomical Polarimetry: Current Status and Future Directions [ASP Conf. Ser. 343],* A. J. Adamson et al., Eds. San Francisco: ASP, pp. 171-175, 2005.

[29] A. Baran, "Foreground detection with the DRAO synthesis telescope: methods and models to measure the polarized cosmic microwave background," M.Sc. thesis, University of Manitoba, 2010.

[30] A. Rogers and J. D. Fiege, “Gravitational lens modeling with genetic algorithms and particle swarm optimizers,” *Ap. J.*, 727 (2), pp. 80-98, 2011.

[31] P. S. Potrebko, et al., "A simple geometric algorithm to predict optimal starting gantry angles using equiangular-spaced beams for intensity modulated radiation therapy of prostate cancer," *Med. Phys.*, 34 (10), pp. 3951-3961, 2007.

[32] A. Ahnesjo, "Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media," *Med. Phys.*,16 (4), pp. 577-92, 1989.

[33] A. Niemierko, "Reporting and analyzing dose distributions: a concept of equivalent uniform dose," *Med. Phys.*, 24 (1), pp. 103-10, 1997.

[34] T. Knoos, I. Kristensen, and P. Nilsson, "Volumetric and dosimetric evaluation of radiation treatment plans: radiation conformity index," *Int. J. Radiat. Oncol. Biol. Phys.*, 42 (5), pp. 1169-76, 1998.

[35] J. Stein, et al., "Number and orientations of beams in intensity-modulated radiation treatments," *Med. Phys.*, 24 (2), pp. 149-60, 1997.

3 INVESTIGATION OF FLUENCE PARAMETERIZATIONS FOR PARETO MULTIOBJECTIVE IMRT TREATMENT PLANNING SOFTWARE

An abbreviated version of this chapter will provide the basis for a manuscript entitled "Investigation of Fluence Parameterizations for PARETO Multiobjective IMRT Treatment Planning Software" by Heather Champion*, Jason Fiege, Boyd McCurdy, Peter Potrebko, and Andrew Cull to be submitted to Medical Physics in 2012.

*This author has generated the results, written the text, and developed much of the methodology for this chapter.

3.1 Introduction

Radiation therapy involves inherent tradeoffs between delivering homogeneous prescription doses to the planning target volumes (PTVs) and sparing organs-at-risk (OARs). Intensity-modulated radiation therapy (IMRT) treatment planning systems that are in use today collapse the multiple clinical objectives into a single objective function, so that a human operator must manually adjust weighting coefficients and dose objectives in order to guide the optimization towards a single plan that best represents their preferences in the compromises to be made. Furthermore, beam angles must be iteratively adjusted and the optimization restarted for each configuration if no automatic beam angle optimization software is used. Not only is this approach time-consuming, but there is no guarantee that the resulting plan is Pareto-optimal, where no one objective can be improved without degrading at least one other. While some efforts have been made to use multiobjective optimization, most of these have dealt with fluence map optimization (FMO) alone [1–9], or have optimized beam orientations while treating FMO as a separate “inner loop” problem [10, 11]. Beam angle optimization (BAO) is known to be a difficult non-convex problem [12], requiring a stochastic optimizer such as a genetic algorithm (GA).

Chapter 2 introduced PARETO, a novel treatment planning system that employs a powerful evolutionary optimizer called Ferret to handle the problem of beam angle and beam fluence optimization simultaneously [13]. For this monolithic approach, fluence parameters are not optimized in an inner loop; rather all parameters are treated equally.

The advantage of a global multiobjective approach is that a set of non-dominated solutions is discovered that may approximate the Pareto front in an efficient, unbiased manner, eliminating the need for a human operator to iteratively drive the optimization. This database of non-dominated solutions represents the optimal set of tradeoffs for a particular patient. With PARETO, a radiation oncologist may explore the patient database using a custom designed visualization tool, and then make a more efficient and informed choice of an optimal plan. In this way, PARETO promises to directly improve the accuracy and efficiency of treatment planning.

Previous efforts have been made to validate PARETO by comparing it to a commercial treatment planning system (Eclipse v8.6, Varian Medical Systems, Palo Alto, CA), and to test PARETO's objective functions on a simple cylindrical phantom with three OARs [13]. Preliminary tests also show that PARETO has the ability to optimize the number of beams. As the next major area of research, this work aims to determine the best parameterization of fluence modulation for PARETO, as judged by the quality of the plans produced and the run time.

Beamlet and aperture-based techniques are standard approaches to IMRT inverse planning. Typical beamlet-based techniques divide the beam into a set of elements called 'bixels' that deposit energy in voxels along their path. The amplitude (fluence) of each bixel is treated as a separate parameter in the optimization. Objective functions are formulated based on the total dose deposited by all bixels in the voxels of a region of interest (ROI) [14]. The optimized fluence of a beam represents an ideal intensity map that must be converted into multileaf collimator (MLC) leaf sequences. While beamlet-based approaches are quite useful in allowing flexible fluence modulation and improve the solution quality (e.g. dose conformality) compared to 3D-CRT, the plans can involve noise and artifacts that produce sharp fluence fluctuations unless appropriate smoothing procedures are used. Sharp fluence fluctuations and complex fluence patterns increase the number of monitor units required for delivery (and thus the beam-on time), reduce dosimetric accuracy, and make the post-optimization conversion to MLC leaf velocity trajectories less efficient [15–17].

Aperture-based techniques make use of the physical delivery constraints of the MLC directly in the optimization while cleverly optimizing segment weights defined by important contours or leaf positions [14]. Optimization based on anatomical contours has the advantage of efficiently producing a high quality plan by outlining the projection of the PTV in the beam's-eye-view (BEV) plane, and sparing the OARs by adding additional segments. However, in cases where a concave target wraps partially around an OAR such as the spinal cord, more PTV segments must be added that are increased in fluence near the OAR in order to produce a homogeneous target dose distribution [18]. Thus, although anatomical contours may be used to quickly generate a plan that conforms the dose to the PTV and adequately spares the OARs, further fluence gradients are often necessary in order to produce a uniform target dose.

Kestin *et al*. have developed an alternative approach for selecting aperture shapes by using 3D isodose surfaces calculated from a plan with equally weighted open tangential fields [19]. The isodose surfaces are projected onto the 2D BEV planes. Segments are then defined by the isodose projections, the open fields, and OAR blocked fields. This approach is a simple parameterization that uses physically relevant contours to produce high quality plans.

Direct Aperture Optimization (DAO) is another popular approach for segment-based IMRT [20, 21]. With DAO, both the shapes and weights of the apertures are directly optimized. Stochastic algorithms such as simulated annealing are necessary since aperture shape optimization is a non-convex problem [14]. Thus, small changes in leaf positions are accepted or rejected based on changes in the objective function, the simulated annealing cooling schedule, the physical MLC constraints, and the minimum aperture size and weight that are specified. DAO has the advantage of using a smaller number of segments to deliver highly conformal plans [21].

While all of these approaches to fluence optimization offer certain advantages, none of them are perfectly suited for PARETO. For instance, although beamlet-based techniques offer complete flexibility in local control over any region of a fluence map, a very large population size would be required to fully explore the enormous parameter space, resulting in clinically unfeasible run times [22]. For this reason, we have devised more efficient parameterizations that stem from some of the concepts encountered in beamlet and aperture-based techniques, such as the delineation of critical structures or isodose surfaces on the BEV, the use of simple fluence gradients, and the formation of beamlet groups. We perform an evaluation of the merits of each method using several patient geometries and establish the best approach for future work.

3.2 Methods and Materials

3.2.1 Fitness Functions

In this chapter, we have employed two formulations of the clinical objectives presented in Section 2.2.2. The first objective function is designed to suppress external hot spots and produce a dose distribution that is conformal to the PTV:

$$F_{PTV} = \log_{10}\left\{\frac{1}{2}\left(1 - \sum_{i=1}^{N}\frac{Q_i \cdot (D_i - \bar{D}) \cdot (M_i - \bar{M})}{\sigma_D \sigma_M}\right)\right\}. \tag{3.1}$$

It is based on the statistical correlation between the total dose distribution *D* (standard deviation σ_D) and a logical geometric mask *M* of the PTV that is restricted to a "hot zone" (standard deviation σ_M). A pre-determined weight matrix Q_i specifies the relative weight applied to the interior and exterior of the PTV. The OAR dose-volume histogram (DVH) objective function considers the mean volume-weighted dose of an OAR with a dose threshold of D_0:

$$F_{OAR} = \frac{1}{V_{OAR}}\left[\sum_{i=1}^{N} f_i^p \, \Delta V_{OAR,i}\right]^{1/p}, \tag{3.2}$$

where

$$f_i = \begin{cases} 0, & D_i \leq D_0 \\ D_i/D_0 & D_i > D_0 \end{cases}. \tag{3.3}$$

The variable $\Delta V_{OAR,i}$ represents the volume increment of the i'th bin of the DVH curve, and V_{OAR} is the total volume of the OAR. In this formulation, voxels that receive higher dose are more strongly suppressed and are controlled by the exponent p.

3.2.2 Fluence Modulation

Since PARETO handles a large-scale monolithic optimization that treats beam angles and fluence parameters equally, it is desirable to reduce the number of parameters where possible. In conventional beamlet-based optimization, the fluence amplitude of each ray is treated as a separate parameter [14]. However, as discussed above, this enormous search space is not practical for PARETO. Therefore, as described in the sections that follow, we have devised several fluence parameterizations that drastically reduce the size of the search space. Since beamlet-based approaches optimize every pixel of the fluence map independently, they have the advantage of being able to produce completely arbitrary fluence patterns. Thus, in order to mimic the quality of solutions achieved with beamlet-based optimization, we seek a parameterization that offers a large amount of flexibility in generating various fluence patterns using a small number of parameters. However, "flexibility" is not an easily quantifiable characteristic. Instead, in the sections that follow we use fluence map *complexity* as a proxy for flexibility, since there are convenient metrics for evaluating image complexity that already exist (see Section 3.2.3). Also, note that although a high degree of fluence complexity implies a more cumbersome delivery, in general the greater fluence complexity possible with IMRT is known to improve the quality of treatment plans compared to 3D-CRT [15–17].

Hence, in this work, we seek a parameterization that is capable of producing a high degree of fluence complexity.

For each method described below, rays are only traced if they intersect a mask defined in the BEV that includes the PTV projection modified by a margin parameter. This decreases the computation time significantly and improves PTV conformity (see Section 3.2.2.7). The amplitude of each ray is modified according to the pattern generated by the fluence parameterization. OAR projections in the BEV that overlay the mask are reduced in fluence, as described in Section 3.2.2.6 below. This approach utilizes anatomical contours but also incorporates a greater degree of fluence complexity. Fluence is represented in dimensionless (i.e. relative) units, but the mean PTV dose of any solution may be scaled to any desired mean dose. Thus, fluence maps may range from a minimum value of zero to a maximum value of one. Including the margin parameter, the total number of parameters required for the coplanar runs done in this study is $N_{par} = N_{beam}(N_{PTV} + N_{OAR} + 2)$, where N_{beam} is the number of beams defined, and N_{PTV} and N_{OAR} are the number of parameters used for modulation over the PTV and OAR projections, respectively. For a complete listing of parameters and their search ranges, see Appendix A.

3.2.2.1 Basic Weight Method

For the simplest parameterization, a constant weight (w) for a given beam determines the amplitudes of the rays that lie within the PTV mask in the BEV. Thus, for the basic weight method, only one parameter is required per beam for PTV modulation. The amplitudes of rays that lie within OAR projections in the BEV are modified by reduction factors (see Section 3.2.2.6). This method is the most similar to anatomy-based contour optimization [14], but is adapted for non-segment-based optimization.

3.2.2.2 Linear Gradient Method

Another simple parameterization involves applying multiplied gradients over the PTV mask in the BEV. Each individual gradient forms an intensity map that is described by:

$$g(x_i, y_j) = \frac{1}{2} + \tanh^{-1} g_0 + \tan\theta_x \cdot \frac{(x_i - x_0)}{L} + \tan\theta_y \cdot \frac{(y_j - y_0)}{L}. \qquad (3.4)$$

Here (x_i, y_j) are the coordinates of a pixel on the BEV and (x_0, y_0) are the coordinates of the PTV's isocentre. L is a distance scale representing the size of the BEV. The parameters θ_x and θ_y specify the gradient strength in each direction and are both constrained such that $-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$. The total fluence offset must be allowed to vary between $-\infty$ and $+\infty$ in order to ensure that steep (even vertical) gradients are possible near the edges of the map (see Fig. 3.1). Thus, the inverse hyperbolic tangent function is used as a means of mapping a finite domain ($-1 \leq g_0 \leq 1$) to an infinite range, which is necessary since Ferret is a bounded optimizer. However, in order to constrain the total fluence map to a finite range, we truncate each gradient map between zero and one such that

$$I(x_i, y_j) = \begin{cases} g(x_i, y_j), & 0 \leq g(x_i, y_j) \leq 1 \\ 0, & g(x_i, y_j) < 0 \\ 1, & g(x_i, y_j) > 1 \end{cases} . \tag{3.5}$$

We have also added a fluence constant of 0.5 in Equation 3.4 so that each gradient is centred in the truncation range.

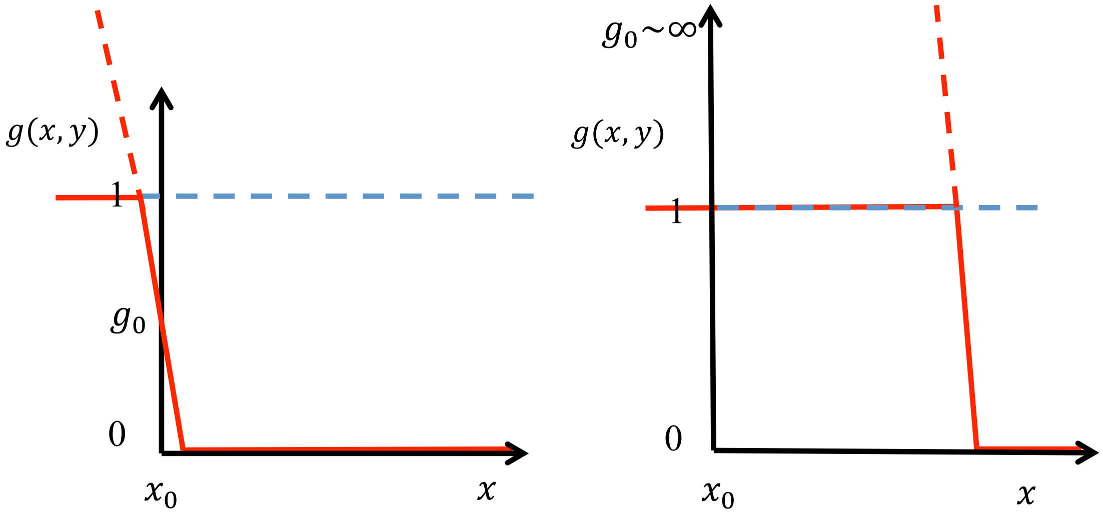

Figure 3.1: (Left) A steep gradient near the PTV isocentre (x_0) requires a small fluence offset, g_0. (Right) A steep gradient near the edge of the map requires a very large fluence offset. If the gradient is vertical, $g_0 \rightarrow \infty$. The solid red curves show the fluence gradients after truncation.

An earlier version of the linear gradient method, called the 'coupled' method, was used with a single gradient per beam for our first tests on a simple cylindrical phantom (see Section 2.2.2; Equation 2.1 is a simplified representation of this method). This method uses a polar rotation angle φ and a scaling parameter θ to specify gradient strength, which implies an unnecessary coupling of the parameters since the gradient strength in each direction depends on two parameters rather than one. Also, the offset was originally specified by a constant bound within the range [-1, 1], such that steep off-centre gradients were difficult to attain. Still, the modulation was found to be sufficient when compared to Eclipse. That is, the dose characteristics of PARETO solutions (as judged by DVHs) were very similar to Eclipse solutions (see Section 2.3.1). Also, the fluence maps of PARETO solutions were qualitatively similar to the Eclipse solutions. Thus, we re-formulate this parameterization using the inverse hyperbolic tangent offset function:

$$g(x_i, y_j) = \frac{1}{2} + \tanh^{-1} g_0 + \tan\theta \cdot \left(\cos\varphi \cdot \frac{(x_i - x_0)}{L} + \sin\varphi \cdot \frac{(y_j - y_0)}{L}\right). \quad (3.6)$$

The fluence map is again truncated using Equations 3.5. We compare this parameterization to the 'decoupled' method (Equation 3.4) in Section 3.3.2. Note that arbitrary gradients may be constructed with both the coupled and decoupled parameterizations, and thus they are mathematically equivalent. However, differences are possible in terms of the optimizer's ability to search the space.

Our first approach to combine several gradients on the fluence map of a given beam was to average the gradient maps using either the mean or median. However, both of these modes have the unwanted effect of decreasing the amount of modulation in the combined map as the number of gradients is increased. Using the mean fluence, gradients are more likely to compensate for each other. Using the median fluence, values closer to the centre of the fluence range are more likely to be chosen.

A better approach for combining individual gradients is to multiply them. While multiplying darkens the resulting image, the relative modulation should generally increase with the number of gradients. We consider three possible implementations. First, it is possible to multiply the gradient maps without truncating them or the resultant image. However, this tends to create maps with hot and cold spots on the edges of the map and relatively little modulation in the middle, especially in the case of a large number of gradients. This is because each linear gradient *must* have its maximum and minimum value on the boundary of the map if it is not truncated. If truncation is applied after the gradients are multiplied, the resultant image tends to show sharp edges on the outer regions of the map where the maxima and minima of the multiplied gradients have been truncated. Thus, we have chosen to truncate each gradient prior to multiplication. Nonetheless, this approach has the disadvantage that in the limit of a large number of gradients, the combined image is very likely to be flat and dark (zero) since the chance that each pixel is multiplied by the lower truncation value increases.

For the linear gradient method, the number of parameters required per beam for the PTV fluence pattern is

$$N_{PTV} = 3N_{grad}, \tag{3.7}$$

where N_{grad} is the number of gradients. Thus, multiplying linear gradients produces greater fluence complexity than is possible with the basic weight method, and still uses a small number of parameters compared to beamlet-based optimization.

3.2.2.3 Discrete Cosine Transform Method

Another method to produce a relatively complex intensity pattern using a small number of parameters is to take an image transform of a few low-frequency parameters. The inverse Fourier transform of a real or complex-valued function does not generally produce a real image. Therefore, the 2D inverse discrete cosine transform (DCT) is more appropriate since it uses real frequency components to produce a real image [23]:

$$g(x_i, y_j) = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \alpha_m \alpha_n C(u_m, v_n) \cos\frac{(2x_i+1)u_m\pi}{2N} \cos\frac{(2y_j+1)v_n\pi}{2N}, \quad (3.8)$$

where

$$\alpha_m = \begin{cases} \sqrt{\frac{1}{N}}, & u_m = 0 \\ \sqrt{\frac{2}{N}}, & u_m = 1,2,\dots,N-1 \end{cases}, \quad (3.9)$$

and N is the size of the map. In PARETO, the frequency map C is padded with zeros in order to avoid aliasing artifacts, such that N is chosen to be twice the size of the BEV.

In order to first see whether the GA is capable of producing a variety of functions in the spatial domain using a few low-frequency parameters, we perform independent 1D tests (completely removed from PARETO). These 1D tests are analogous to the 2D FMO problem handled by PARETO. We employ a 1D linear gradient function of slope m,

$$y(x_i) = mx_i, \quad 0 \le x_i \le 60, \quad (3.10)$$

and a step function,

$$y(x_i) = \begin{cases} 0.5, & 20 \le x_i \le 40 \\ 0, & \begin{matrix} 0 \le x_i < 20 \\ 40 < x_i \le 60 \end{matrix} \end{cases}, \quad (3.11)$$

to describe two different desired fluence patterns. The 1D frequency maps are described by:

$$C(u_m) = \begin{cases} k & u_m \le \rho \\ 0 & u_m > \rho \end{cases}, \quad (3.12)$$

where ρ determines the number of frequency parameters (k) included in the optimization.

We choose three values of ρ (9, 19, and 59) for three independent tests with 10, 20, and 60 parameters, respectively. The objective function is specified by the residual sum of squares,

$$F = \sum_{i=1}^{N}(y(x_i) - g(x_i))^2, \tag{3.13}$$

where $g(x_i)$ is obtained using Equation 3.12 in the inverse DCT (Equation 3.8), and $y(x_i)$ is the discrete data of either Equation 3.10 or 3.11. Figure 3.2c shows that Ferret has no difficulty reproducing a simple linear gradient function using as little as 10 frequency parameters ($F = 7.76\times10^{-4}$). However, for the step function, the optimal solution is excellent using all 60 frequencies ($F = 1.75\times10^{-5}$; see Fig. 3.2d) but becomes increasingly worse with lower numbers of frequencies ($F = 0.312$ for 10 frequency parameters; see Fig. 3.2e, f). Still, these tests show that searching the DCT space is feasible for FMO.

To implement the cosine transform method in PARETO, nonzero frequency amplitudes are defined on a 2D frequency map C within a certain radius ρ from the origin (which represents the DC component, located at the top-left of the map):

$$C(u_m, v_n) = \begin{cases} k, & \sqrt{{u_m}^2 + {v_n}^2} \leq \rho \\ 0, & \sqrt{{u_m}^2 + {v_n}^2} > \rho \end{cases}. \tag{3.14}$$

The transformed image (Equation 3.8) is used for modulation over the PTV mask in the BEV (the centre of the image is registered to the isocentre of the PTV). Increasing ρ includes higher nonzero frequency components, resulting in greater modulation in the spatial domain. Thus, the number of parameters required per beam for PTV modulation is equal to the number of pixels that lie within the radius ρ from the origin. This number can be approximated by the area of the quadrant defined by the origin and two radii of length ρ on the boundary of the map:

$$N_{PTV} \sim \frac{1}{4}\pi\rho^2. \tag{3.15}$$

For the exact numbers of PTV fluence parameters used in this work, see Table 3.

In our first version of the cosine transform method we specified the range of the nonzero frequency parameters k such that pixel values in the spatial domain were constrained within the range [-0.5, 0.5]. A constant value of 0.5 was added to the fluence map to ensure that its range was [0, 1]. To constrain the frequencies, the maximum value of the inverse DCT of Equation 3.14 was used (which occurs when all frequency parameters k are simultaneously equal to one):

$$-\frac{0.5}{\max\left(g(x_i, y_j)\right)} \leq k \leq \frac{0.5}{\max\left(g(x_i, y_j)\right)}\,. \tag{3.16}$$

However, this approach had the disadvantage that as ρ increased, the typical fluence range decreased (see Figure 3.3). Thus, to obtain the normalization coefficient, we instead choose to use a large number (10^3) of random transforms to find a typical maximum pixel value ($\bar{g}$) in the spatial domain. To find $\bar{g}$, transforms are calculated with Equation 3.8 where

$$C(u_m, v_n) = \begin{cases} z_{m,n}, & \sqrt{{u_m}^2 + {v_n}^2} \leq \rho \\ 0, & \sqrt{{u_m}^2 + {v_n}^2} > \rho \end{cases}, \tag{3.17}$$

and $z_{u,v}$ is a random number between -1 and 1. Thus,

$$-\frac{0.5}{\bar{g}} \leq k \leq \frac{0.5}{\bar{g}}\,. \tag{3.18}$$

To produce the fluence map, the frequency parameters are constrained within this range and a constant value of 0.5 is added to Equation 3.8 in order to centre the maps within the truncation range specified by Equation 3.5 (see Figure 3.4).

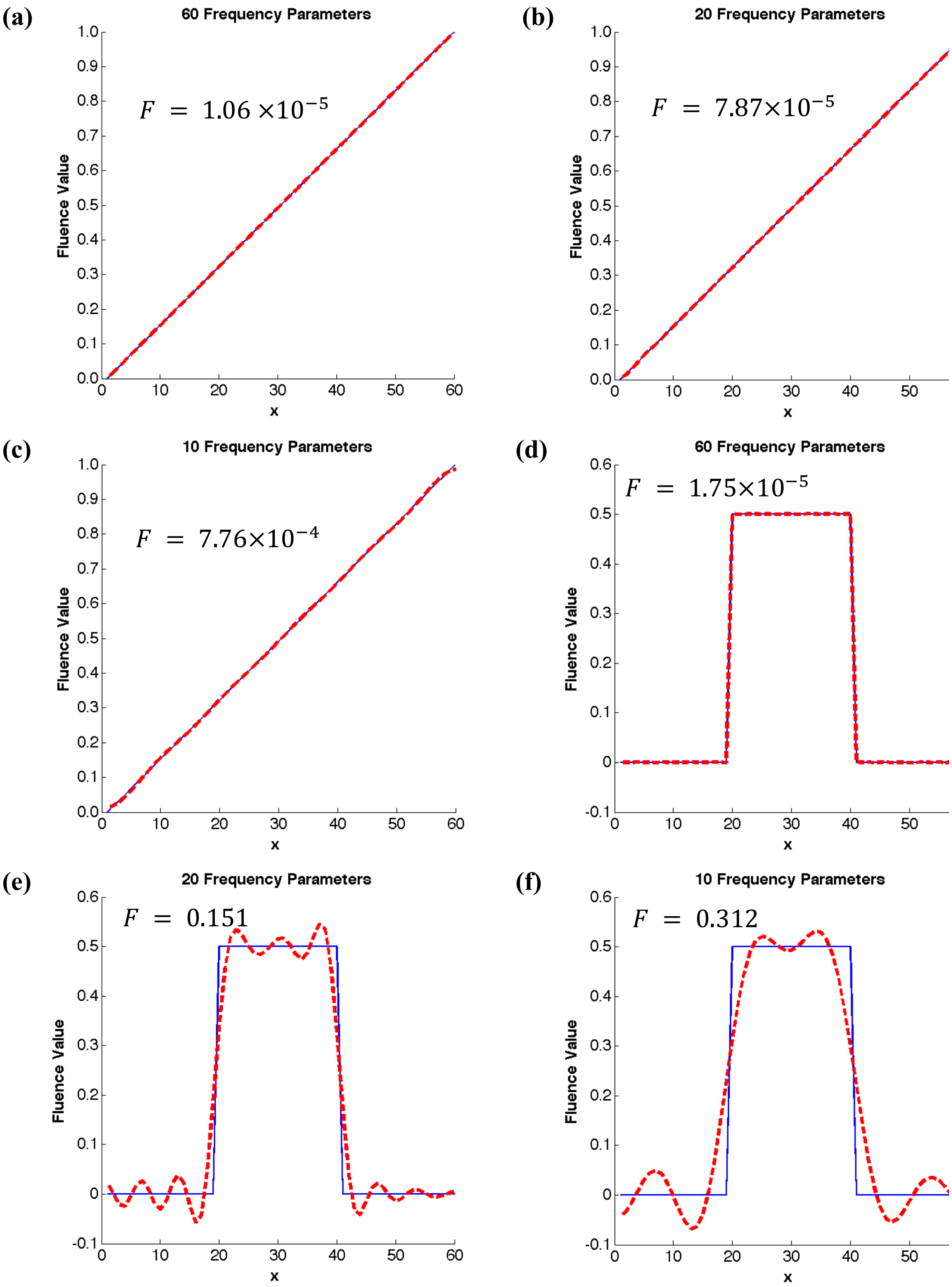

Figure 3.2: An independent test showing the model (dashed red) that Ferret was able to generate for the data (blue) using the inverse DCT. (a) All 60 frequencies of the frequency map were optimized by Ferret to generate a simple gradient. (b) Only 20 non-zero low-frequency parameters were given to Ferret to control. (c) Only 10 non-zero low-frequency parameters were used. (d) All 60 frequencies were used to model a step function. (e) Only 20 low-frequency parameters were used. (f) Only 10 low-frequency parameters were used.

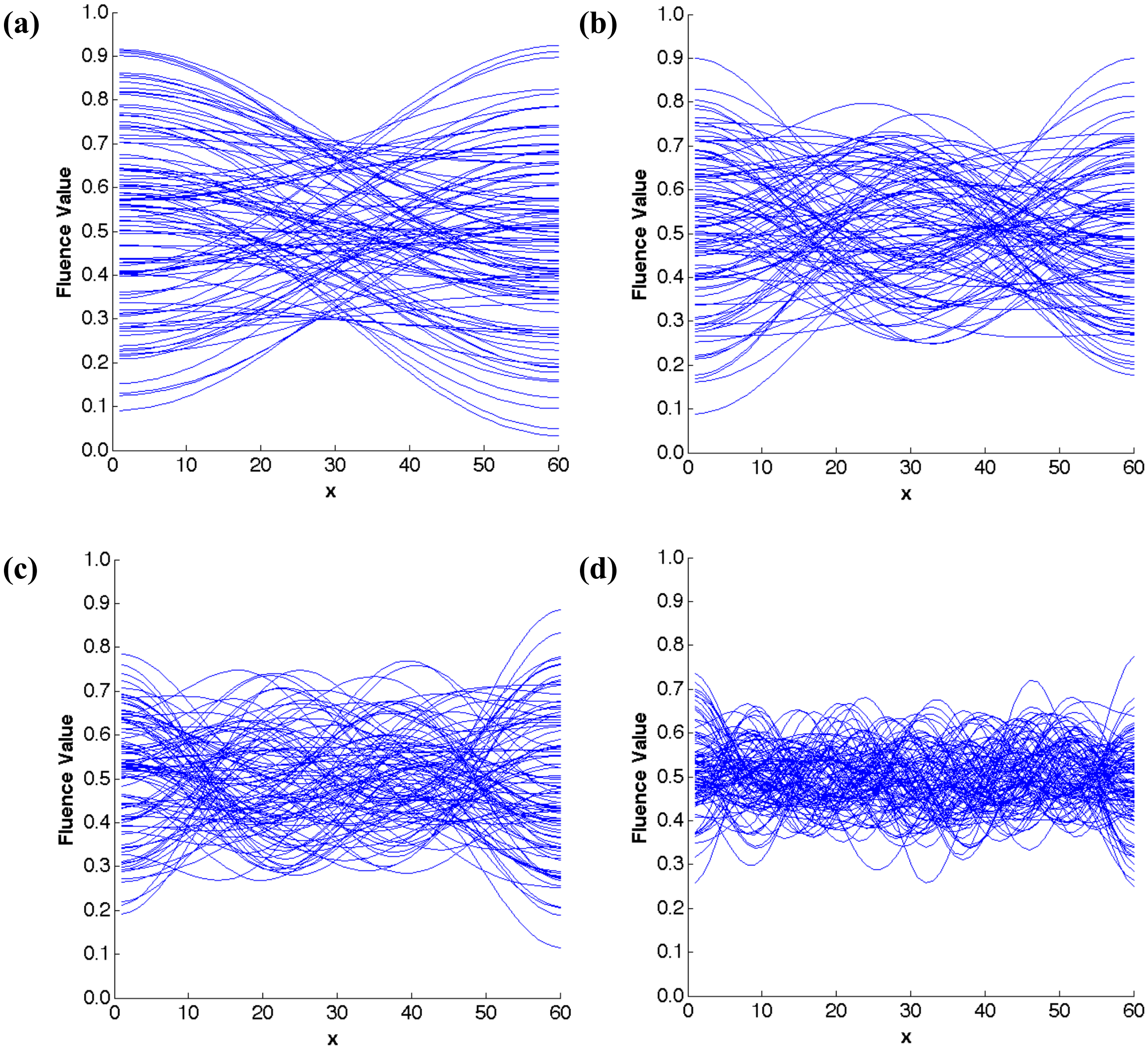

Figure 3.3: A large number of randomly generated 1D inverse DCTs using a frequency range set by Equation 3.16, with a constant value of 0.5 added in the spatial domain. The radius of the nonzero frequency components is (a) $\rho = 1$, (b) $\rho = 2$, (c) $\rho = 3$, and (d) $\rho = 8$.

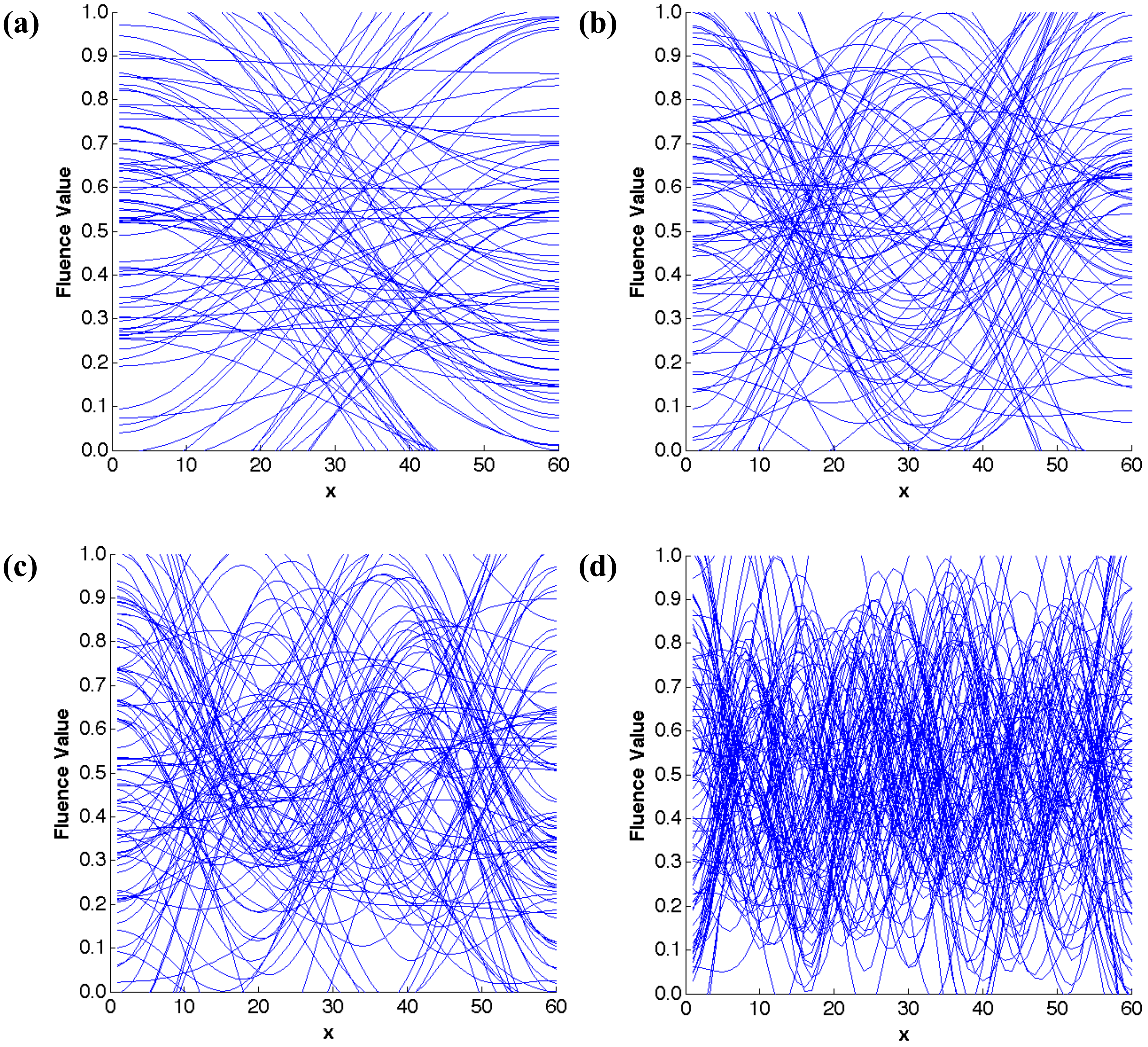

Figure 3.4: A large number of randomly generated 1D inverse DCTs using a frequency range set by Equation 3.18 and truncation according to Equation 3.5 (with a constant value of 0.5 added prior to truncation). The radius of the nonzero frequency components is (a) $\rho = 1$, (b) $\rho = 2$, (c) $\rho = 3$, and (d) $\rho = 8$.

3.2.2.4 Beam Group Method

Beamlet-based optimization is a flexible way of optimizing the dose deposited by each ray traced. However, as discussed previously, it is desirable to limit the number of fluence parameters used in PARETO. A simple variant of beamlet-based optimization is to divide the rays into a coarse grid of square 'beam groups'. Rather than using the whole BEV (the size of which is determined by the maximum extent of the PTV in any possible view), the beam groups are defined on a square grid with side length equal to the maximum dimension in x or y of the mask of the PTV and its margin. This approach avoids wasting many beam groups on the exterior of the mask. The central intensity of each beam group is treated as a free parameter, while the amplitudes of the individual rays are obtained by bilinear interpolation over the beam groups, such that

$$g(x_i, y_j) = \frac{h_{m,n}}{(x_{m+1}-x_m)(y_{n+1}-y_n)}(x_{m+1} - x_i)(y_{n+1} - y_j) + \frac{h_{m+1,n}}{(x_{m+1}-x_m)(y_{n+1}-y_n)}(x_i - x_m)(y_{n+1} - y_j) + \frac{h_{m,n+1}}{(x_{m+1}-x_m)(y_{n+1}-y_n)}(x_{m+1} - x_i)(y_j - y_n) + \frac{h_{m+1,n+1}}{(x_{m+1}-x_m)(y_{n+1}-y_n)}(x_i - x_m)(y_j - y_n)\,, \tag{3.19}$$

where $(x_m, y_n), (x_{m+1}, y_n), (x_m, y_{n+1}), (x_{m+1}, y_{n+1})$ are coordinates of the four closest neighbours of (x_i, y_j) on the coarse grid with intensities $h_{m,n}, h_{m+1,n}, h_{m,n+1}, h_{m+1,n+1}$. The range of h was chosen so that $-1 \leq h \leq 2$, and Equation 3.5 was used to truncate the interpolated intensity map in order to allow sharper gradients. The number of PTV fluence parameters required is

$$N_{PTV} = (N_{coarse})^2, \tag{3.20}$$

where N_{coarse} is the side length of the coarse intensity map h. For the runs discussed in Section 3.3.3.3, coarse grid sizes of 3 by 3, 5 by 5, 7 by 7, and 9 by 9 were used. Thus, this method allows the GA great flexibly in sculpting the dose but still reduces the size of the search space substantially compared to beamlet-based optimization.

3.2.2.5 Isodose-based Projection Method

This parameterization is derived from the approach used by Kestin *et al.* [19] for isodose-based aperture optimization. First, a dose distribution is obtained for a particular parameter set by equally weighting non-modulated fields at the given beam angles. For this calculation, each field conforms to the mask of the projection of the PTV in the BEV modified by the margin parameter (see Section 3.2.2.7). From this dose distribution, several dose levels are selected uniformly between the minimum dose in the PTV voxels and 95% of the maximum dose in the PTV voxels. For each dose level, the faces and vertices of a 3D isodose surface are calculated for the PTV voxels using MATLAB's built-in "isosurface" function. This function connects points with the same dose level in the volume data. The image of the projection of each isosurface (D_n) on the BEV forms a 2D mask, M_n:

$$M_n = \begin{cases} 1, & \mathrm{proj}_{\mathrm{BEV}}\big(D_n(x', y', z')\big) > 0 \\ 0, & \mathrm{proj}_{\mathrm{BEV}}\big(D_n(x', y', z')\big) = 0 \end{cases}. \tag{3.21}$$

The sum of all masks produces an indexed image $g(x_i, y_j)$ that is used to define regions on the BEV according to each label m ($m > 0$):

$$g(x_i, y_j) = \sum_{n=1}^{N_{iso}} M_n \,, \tag{3.22}$$

and

$$I(x_i, y_j) = \begin{cases} w_m, & g(x_i, y_j) = m \\ 1, & g(x_i, y_j) = 0 \end{cases}. \tag{3.23}$$

The GA controls weight parameters (w_m) for each region. The number of isodose surfaces calculated (N_{iso}) corresponds to the number of PTV fluence parameters that are required. The map $I(x_i, y_j)$ is used to modulate the amplitude of rays already traced through the treatment volume, and the rays are re-convolved with the patient scatter kernel. We have chosen the range of $I(x_i, y_j)$ such that the rays may be increased or decreased in amplitude ($0 \leq w_m \leq 2$). Thus, the *effective* fluence map may be considered as the product of the initial intensity value used for the ray trace (0.5) and $I(x_i, y_j)$.

This method has the advantage of adjusting the fluence to accommodate for dose inhomogeneities in the target due to a solution's particular beam orientations. This benefit is unique to our implementation, since Kestin *et al*. [19] did not optimize beam orientations. Thus, the strength of this parameterization should be seen on runs done with a smaller number of beams (resulting in greater dose inhomogeneity in the PTV) and on patient geometries that involve greater attenuation across the PTV.

3.2.2.6 OAR Fluence Modulation

For each OAR projection that overlays the mask of the PTV and its margin in each BEV, our standard approach is to use a reduction parameter $0 < q < 1$ that multiplies the PTV fluence. Thus, relatively smooth fluence gradients may be applied over the PTV by one of the above methods, while sharp discontinuities result at OAR boundaries due to the reduction parameters. This approach was used for initial development, and continues to provide efficient shaping of the fluence around critical structures. However, we have tested the effect of omitting the reduction parameters in conjunction with the beam group method (see Section 3.3.3.3). We have also tested the effect of multiplying a single linear gradient function by the PTV fluence where each OAR overlaps the PTV mask. Equations 3.4 and 3.5 specify the gradient function for each OAR.

3.2.2.7 Margin Parameter

The length of a PARETO run is influenced by the time required to evaluate the objective function and also by the population size. To reduce the former, we have found it effective to reduce the number of rays traced by specifying a margin parameter for each BEV. In this approach, the GA selects an optimal fluence boundary by dilating or eroding the projection of the PTV, and only rays that lie inside it need to be traced. Note that in no way does this modify the contours defining the PTV that include a clinical margin added for patient motion and setup uncertainties (see Section 1.2). Rather, this is a fluence map optimization technique designed to identify the optimal fluence boundary. A circular dilation kernel is used and a free parameter determines the radius of the kernel, r_{kernel}, and thus the number of pixels eroded or dilated:

$$\text{if } r_{kernel} < 0, \text{ then erode by } r_{kernel} \text{ pixels.} \tag{3.24}$$

$$\text{if } r_{kernel} > 0, \text{ then dilate by } r_{kernel} \text{ pixels.} \tag{3.25}$$

The range of the margin parameter is chosen such that

$$-11 \leq r_{kernel} \leq 11, \tag{3.26}$$

where r_{kernel} is an integer. Alternatively, the edge could be taken as the exact outline of the PTV, or our simple gradient methods could be used to find an optimal edge. These approaches have the disadvantages of limited flexibility, and poor PTV dose conformity, respectively.

3.2.3 Methodology for Comparison of Fluence Parameterizations

In order to determine the best fluence parameterization for PARETO, several evaluation criteria must be considered. First, an efficient parameterization is desirable. By ‘efficient’, we mean a parameterization that results in good solutions in a small run time. Parameterizations with a very large number of parameters (such as beamlet-based approaches) may significantly increase the population size needed for the GA to fully explore the space, and thus the total number of evaluations of the objective function. However, if the population size is kept constant, a change in the number of parameters does not significantly affect the run time (see Appendix B, Fig. B.1a) except for parameterizations where increasing the number of parameters directly increases the number of computations required to evaluate the objective function (e.g. 3D isosurface calculations). In general, the type of computation involved also affects the run time (e.g. image transformation, interpolation, 3D isosurface calculation). Thus, to compare the efficiency of different fluence parameterizations, we select a run from each method that results in high solution quality (see discussion below) and plot its run time (see Fig. 3.47). At present, runs generally require up to a day depending on the size of the data set (see Appendix B, Fig. B.1c), the number of CPUs used, the population size, and the number of generations. However, in future work we will optimize the code for speed and evaluate PARETO's GPU-based ray tracer in order to reduce this to clinically viable run times.

Perhaps the most relevant way to evaluate a particular parameterization is by the GA's ability to search the parameter space and produce high-quality solutions. Ideally, different parameters would control different aspects of the fluence model, resulting in an easier search. In practice, however, some parameters are correlated. The quality of the solutions present in the final non-dominated set acts as a proxy for the searchability of the parameter space. We postulate that clinically desirable solutions are ones that are well balanced, doing well in both PTV conformity and OAR sparing. However, in practice, a human decision maker is required to select particular tradeoffs for a certain patient. The radiation oncologist may favour some objectives over others, depending on the situation. The advantage of a multiobjective approach is to offer the decision maker a set of plans that are equally good according to our multiobjective criteria. Still, if we expect well-balanced solutions to be preferred more frequently, a certain parameterization may be considered superior to another if it typically produces a tradeoff surface with a lower "knee" (where solutions are simultaneously low in all objective function values), or a smaller spread of solutions about the knee. Thus, runs may be visually compared by overlaying non-dominated solution sets in the planes of PTV conformity versus OAR dose. Our previous efforts to compare various runs projected the solutions into the plane of PTV conformity versus quadrature-averaged OAR dose [13]. However, we have found that in some cases this approach can obscure trends that are visible on individual projections of OAR dose. Therefore we avoid quadrature averaging in this work in favour of showing multiple projections of the tradeoff surface.

Although runs may be visually compared relatively easily, a quantitative analysis is more difficult since the knee exists in multi-dimensional objective function space. Therefore, rather than trying to fit a multi-dimensional surface populated by thousands of points to find the location of the knee, we employ a different measure of overall solution quality. Since individual solution quality is encoded by the objective functions (as described in Section 3.2.1), we merge the optimal solutions of each run using Ferret's tournament selection operator, as described in Qubist's User Guide [24]. Tournament selection iteratively compares pairs of solutions, at each step rejecting inferior solutions (see Section 1.3). Thus, only non-dominated solutions remain in the merged set, and the number of solutions from each run that survive the operation can be used to judge the overall quality of one run compared to another. We note that the genes of a particular solution correspond to the parameterization that generated it, thus the number of genes may vary across solutions in the merged set. Also, in order to treat each run fairly such that runs with a small number of good solutions are not overwhelmed by heavily populated sets, we weight the number of "winners" ($N_{win,i}$) of each run by the inverse of the number of solutions that "compete" ($N_{comp,i}$). Thus, for run 'i' we define the survival fraction (SF) as

$$SF_i = \frac{N_{win,i}/N_{comp,i}}{\sum_{i=1}^{M} N_{win,i}/N_{comp,i}}\,, \tag{3.27}$$

where M is the number of runs being used to construct the merged optimal set. While this approach does not prefer well-balanced, "knee" solutions to more extreme solutions, it is preferable from a MOO standpoint since it uses the number of mathematically non-dominated solutions to identify a superior tradeoff surface. Furthermore, since our objective functions include dose thresholds for each OAR, truly extreme solutions that are very high in dose to a certain OAR are not as likely to be present in the final optimal set (see Section 2.2.2). Still, it is possible that a run with a lower knee contains inferior "extreme" solutions than a run with a higher knee, and thus could produce a smaller survival fraction. Therefore, both a visual comparison and a tournament competition are helpful in selecting a superior run.

There are two possible approaches to compare more than two runs. A merged optimal set from the solutions of all runs is sufficient if we only desire to know the ultimate winner of the competition. The run with the greatest survival fraction has the best overall quality of solutions in its database. However, if we desire to rank the runs in order from best to worst, then we must compare the runs in pairs in a "round robin" tournament where each run "plays" every other run. This is due to the fact that it is possible for the best run to completely dominate *all* the solutions of the other runs in the merged optimal set of *all* runs. If this occurs, information regarding the ranking of the other runs would be lost since their solutions would be completely absent in the merged set. Therefore, scoring each run by its performance in the round robin tournament is the best way to rank the runs. The survival fraction (Equation 3.27) is used for each comparison of two runs, and at the end of the round robin the mean survival fraction is used to determine the score of each run.

The survival fraction of solutions in a merged optimal set provides a straightforward way to assess whether one run is superior to another. However, there are other interesting features of a merged optimal set that can be explored by visualizing the set in the planes of PTV versus OAR fitness while colour-coding the solutions according to the particular parameterization that generated them. This provides a clear visualization of localized regions of relative dominance. For instance, the knee of a certain run may completely dominate the knee of a second run, but the second run may be better at producing clusters of non-dominated solutions at the extrema (see Fig. 3.11c, d).

Finally, another measure of the quality of a parameterization is its ability to produce arbitrary fluence patterns. As discussed in Section 3.2.2, we refer to fluence complexity as a proxy for flexibility. For each parameterization, we expect an increase in the number of parameters to increase the fluence complexity. Also, for most patient geometries we expect an increase in fluence complexity to increase the overall quality of the resulting solutions up to a saturation point.

There are several metrics that can be used to evaluate fluence map complexity. First, we consider the modulation index introduced by Webb *et al*. [25] and generalized to 2D by Giorgia *et al*. [15]. The modulation index quantifies the complexity of a fluence map by creating a spectrum of the number of intensity changes between adjacent pixels that exceed a certain fraction of the standard deviation of the map. The motivation for this particular calculation is to relate the magnitude of local pixel variations to the overall characteristic deviation of the beam intensity [25]. Llacer *et al.* developed an alternative metric that also considered local pixel variations using a root sum over neighbouring pixels [26]. However, while it responded well to small or large differences between adjacent pixels, it did not relate these to the amount of variation in the map as a whole.

In our implementation of the modulation index, we only consider pixels that lie within the logical mask of the PTV and its margin in the BEV (M). We first define the functions Δx, Δy, Δxy which contribute to the spectra if the intensity changes in the x direction, y direction, and xy diagonal of the fluence map I are above a certain fraction f of the standard deviation of the map (σ_I):

$$\Delta\mathrm{x}_{i,j}(f) = \begin{cases} 1 & \left|I_{i,j} - I_{i+1,j}\right| > f\sigma_I \ \wedge \ M_{i,j} > 0 \ \wedge \ M_{i+1,j} > 0 \\ 0 & \left|I_{i,j} - I_{i+1,j}\right| \leq f\sigma_I \ \vee \ M_{i,j} = 0 \ \vee \ M_{i+1,j} = 0 \end{cases}, \quad (3.28)$$

$$\Delta\mathrm{y}_{i,j}(f) = \begin{cases} 1 & \left|I_{i,j} - I_{i,j+1}\right| > f\sigma_I \ \wedge \ M_{i,j} > 0 \ \wedge \ M_{i,j+1} > 0 \\ 0 & \left|I_{i,j} - I_{i,j+1}\right| \leq f\sigma_I \ \vee \ M_{i,j} = 0 \ \vee \ M_{i,j+1} = 0 \end{cases}, \quad (3.29)$$

$$\Delta\mathrm{xy}_{i,j}(f) = \begin{cases} 1 & \left|I_{i,j} - I_{i+1,j+1}\right| > f\sigma_I \ \wedge \ M_{i,j} > 0 \ \wedge \ M_{i+1,j+1} > 0 \\ 0 & \left|I_{i,j} - I_{i+1,j+1}\right| \leq f\sigma_I \ \vee \ M_{i,j} = 0 \ \vee \ M_{i+1,j+1} = 0 \end{cases}. \quad (3.30)$$

Next, we define the difference spectra z_x, z_y, and z_{xy}:

$$z_x(f) = \frac{1}{N_x} \sum_{i=1}^{N_x-1} \sum_{j=1}^{N_y} \Delta\mathrm{x}_{i,j}(f), \quad (3.31)$$

$$z_y(f) = \frac{1}{N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y-1} \Delta\mathrm{y}_{i,j}(f), \quad (3.32)$$

$$z_{xy}(f) = \frac{1}{N_{xy}} \sum_{i=1}^{N_x-1} \sum_{j=1}^{N_y-1} \Delta\mathrm{xy}_{i,j}(f), \quad (3.33)$$

where N_x and N_y represent the size of the fluence map in the x and y directions, respectively. In order to arrive at a scalar metric for quantifying fluence complexity, the average difference spectrum,

$$z(f) = \frac{1}{3}\left[z_x(f) + z_y(f) + z_{xy}(f)\right], \quad (3.34)$$

is used, and the modulation index (MI) is defined as a discrete sum over f with a step size of 0.01 ($f_n = 0.01n$):

$$MI(I) = \sum_{n=1}^{10} z(f_n)\ . \tag{3.35}$$

Recently Nauta *et al*. [27] have shown that a fractal dimension analysis of fluence images using a variogram method is superior to the 2D modulation index in its ability to distinguish between moderate and highly modulated fields for head and neck carcinomas. The fractal dimension of a surface increases with the amount of structural detail in an image. An image with a high fractal dimension can be considered a good representation of a mathematical fractal surface, such that there is little variation in the amount of structural detail visible at different scales. Thus, the fractal dimension can be thought of in terms of the similarity of pixel values as a function of the distance between them (the autocorrelation). To define a fractal profile, Nauta *et al.* use the individual rows and columns of the image. The profile has a high fractal dimension if the autocorrelation of a profile decreases gradually as a function of distance. On the other hand, the profile may not be a good representation of a fractal profile if the autocorrelation decreases steeply as a function of distance. Nauta *et al*. employ the semivariogram (which is widely used in spatial statistics) to express the autocorrelation function. The semivariogram models the variance of the random variables $Z(\vec{x} + \vec{h})$ and $Z(\vec{x})$ as a function of the separation between them, called the lag ($\vec{h}$) [28]:

$$\Gamma(\vec{h}) = \tfrac{1}{2}\langle[Z(\vec{x} + \vec{h}) - Z(\vec{x})]^2\rangle = \tfrac{1}{2}\mathrm{Var}[Z(\vec{x} + \vec{h}) - Z(\vec{x})]. \tag{3.36}$$

This assumes that for all $\vec{h}$ and all $\vec{x}$,

$$\langle Z(\vec{x} + \vec{h}) - Z(\vec{x})\rangle = 0. \tag{3.37}$$

Also, the semivariogram is assumed to depend only on the lag and not on the location $\vec{x}$. The empirical semivariogram is calculated by considering pairs of points that are separated by distance h [29]:

$$\gamma(h) = \frac{1}{2}\frac{1}{N}\sum_{i=1}^{N}[z(x_i + h) - z(x_i)]^2, \tag{3.38}$$

where N is the number of pairs of data points separated by lag h. The semivariogram may be expressed in terms of the autocorrelation function $C(h)$ as

$$\begin{aligned}\gamma(h) &= \frac{1}{2}\{\langle z(x+h)^2\rangle - 2\langle z(x)z(x+h)\rangle + \langle z(x)^2\rangle\} \\ &= C(0) - C(h).\end{aligned} \tag{3.39}$$

In order to calculate the fractal dimension, we must obtain an expression for how the semivariogram (or autocorrelation function) behaves as a function of lag for a fractal profile. First, when data is assumed to be fractal, the power spectral density $G(f)$ is a function of the spatial frequency *f*, and goes as

$$G(f) = Cf^{-\alpha} \quad (1 < \alpha < 3), \tag{3.40}$$

where C and α are constants [29]. Using the Wiener-Khintchine relation, we may express the autocorrelation function in terms of the power spectral density function for a fractal (Equation 3.40) [29]:

$$C(h) = \int_0^\infty G(f)\cos 2\pi hf\, df = \int_0^\infty Cf^{-\alpha}\cos 2\pi hf\, df. \tag{3.41}$$

Thus from Equations 3.39 and 3.41, the semivariogram of a fractal profile goes as:

$$\gamma(h) = \int_0^\infty Cf^{-\alpha}(1 - \cos 2\pi hf)df, \tag{3.42}$$

$$\gamma(h) = 2\int_0^\infty Cf^{-\alpha}\sin^2 \pi hf\, df, \tag{3.43}$$

$$\gamma(h) = 2C(\pi h)^{\alpha-1}\int_0^\infty u^{-\alpha}\sin^2 u\, du, \tag{3.44}$$

where $u = \pi hf$. Finally, grouping all constants into constant *V*,

$$\gamma(h) = Vh^{\alpha-1} \quad (1 < \alpha < 3). \tag{3.45}$$

Murata and Saito [29] relate the constant α to the fractal dimension (FD) by

$$FD = \frac{5-\alpha}{2}, \tag{3.46}$$

and so

$$\gamma(h) = Vh^{4-2FD}. \tag{3.47}$$

Thus, to calculate the fractal dimension, Nauta *et al*. use the slope (m) of a plot of $\log\gamma$ versus $\log h$:

$$FD = \frac{4-m}{2}. \tag{3.48}$$

However, the range of lag that obeys the power law (Equation 3.47) extends to less than about 10% of the profile length [29]. Therefore, the particular linear range of each profile (row or column) of the image should be determined on a $\log\gamma$ versus $\log h$ plot (see Figure 3.5). To do this, we choose to fit the data using a least squares fitting algorithm. The coefficient of determination is defined as

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{3.49}$$

where y represents the data set, $\bar{y}$ the mean data value, and f the modeled values. The data point with the largest lag is removed from the set if the R^2 value is less than or equal to 99.9%, and the remaining data is refit. This process is repeated until R^2 is greater than 99.9% (a value chosen to represent a good fit). This automated approach is more accurate than that of Nauta *et al*. who used a constant linear range for all profiles. However, it may not be practical in terms of computation speed. The method may be improved by performing a binary search for the end data point of the linear range that results in a fit where $R^2 \geq 99.9\%$.

To find the fractal dimension of the 2D image, Nauta *et al*. add a constant of one to the mean fractal dimension of all rows and columns of the image (the fractal dimension of a 2D surface lies between 2 and 3). In our implementation, we again consider an image with an irregular boundary as determined by the BEV mask. Thus, the profiles are of varying length in general (see Fig. 3.5).

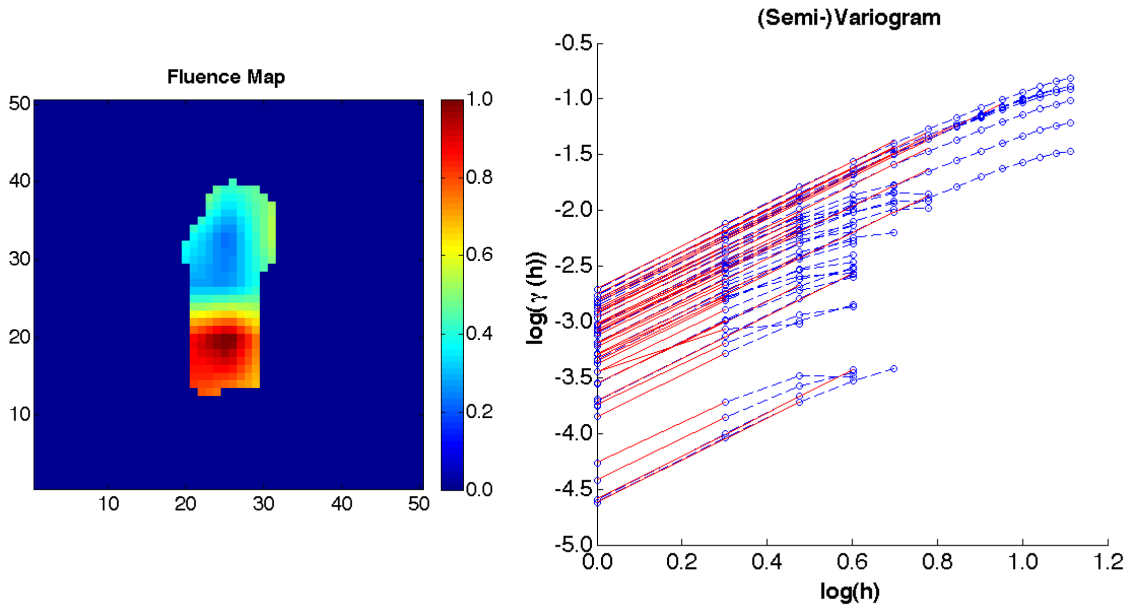

Figure 3.5: Variogram for each profile (row or column) of the PTV mask in the fluence map shown on the left. Red curves indicate linear fits made to each profile.

A final metric that we use to evaluate fluence map complexity is the gradient entropy within the BEV mask. For an intensity distribution I with an irregular boundary, z represents the total set of x and y derivatives:

$$z = \left[\nabla_x I\left(x_i, y_j\right), \quad \nabla_y I\left(x_i, y_j\right)\right]. \tag{3.50}$$

A histogram of bin width Δz with N_n counts in each bin forms the probability mass function $p(z_n)$:

$$p(z_n) = \frac{N_n}{N}, \quad z_n - \frac{\Delta z}{2} < \mathrm{z} < z_n + \frac{\Delta z}{2}. \tag{3.51}$$

The variable N represents the total number of pixels in the image. Thus, the image gradient entropy (GE) is

$$GE(z) = -\sum_{n=1}^{N_{bins}} p(z_n) \log_2 p(z_n), \tag{3.52}$$

where N_{bins} is the number of histogram bins. Figure 3.6 shows the gradients and histogram of the fluence map presented in Figure 3.5.

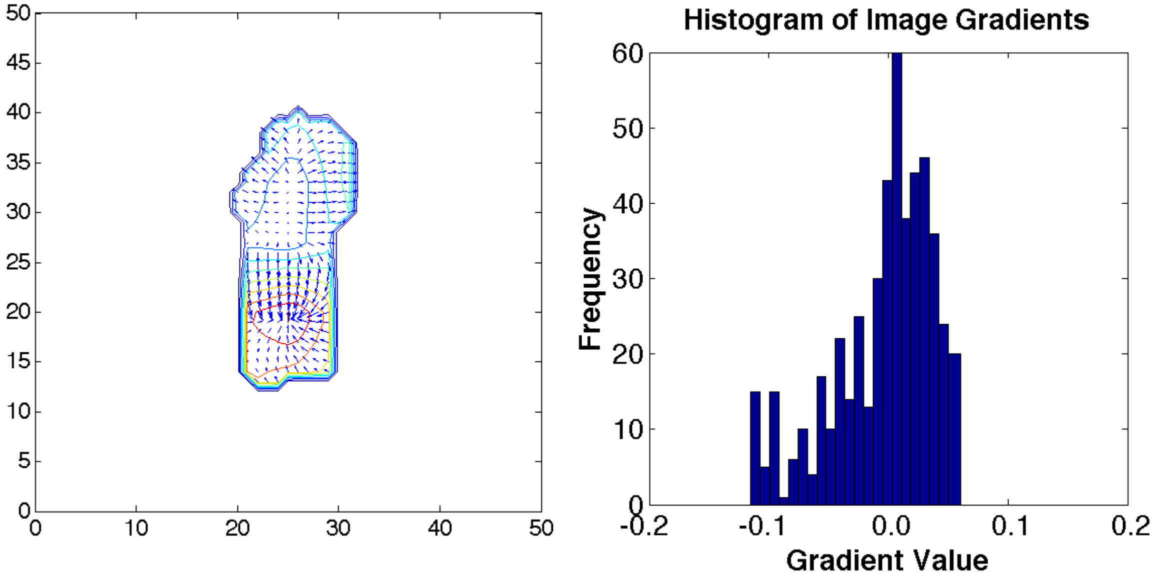

Figure 3.6: (Left) Image gradients (in x and y) of the fluence map shown in Figure 3.5. (Right) Histogram of both x and y gradients.

In order to compare different PARETO runs in terms of fluence complexity, each metric is evaluated on the fluence maps of every optimal plan, and the mean value of all maps is calculated in order to visualize potential trends in fluence complexity (for example, see Fig. 3.24). Also, the mean value of a metric may be calculated for the fluence maps of each optimal solution in order to visualize trends on the tradeoff surfaces (see Fig. 3.42).

3.2.4 Description of Patient Data Sets

We use several data sets representing different geometries to make a full assessment of each fluence parameterization. The first we present here is a simple phantom representing a realistic paraspinal patient. The target is defined as a C-shaped tumour wrapping around the spinal cord, while the OARs include the left and right kidneys and the spinal cord (see Figure 3.7a). The phantom has cylindrical symmetry about the z-axis. Two patient data sets are employed to test more complicated geometries. The first includes a C-shaped tumour wrapped around the spinal cord, with lungs and the spinal cord as OARs (see Figure 3.7b). The second contains a larger tumour wrapped around the spinal cord, with the cauda equina and spinal cord as OARs (see Figure 3.7c).

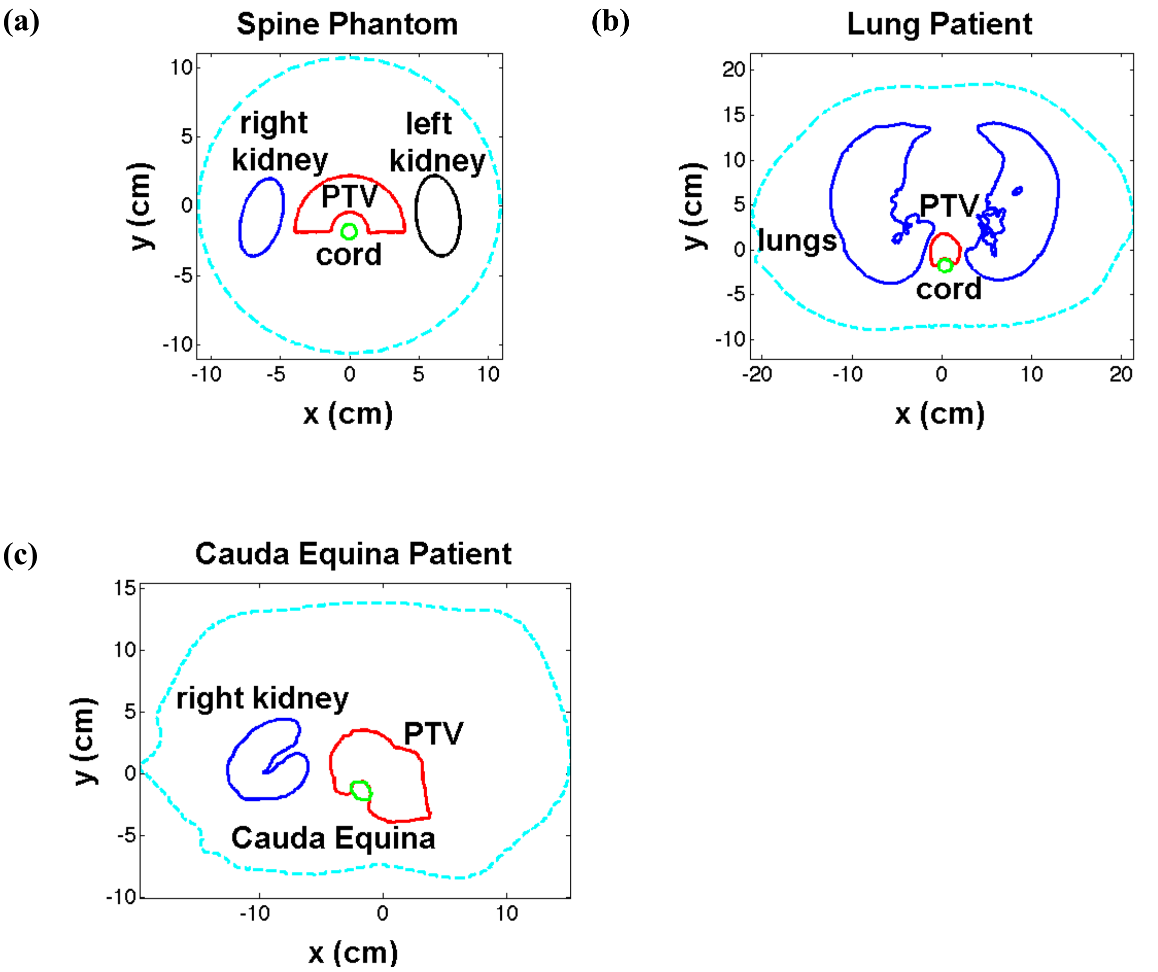

Figure 3.7: (a) A spinal phantom with a C-shaped PTV structure, an OAR representing the spinal cord, and two kidney OARs. The phantom is defined on 13 axial slices in the z direction with the PTV and OARs occupying the central 7 slices. (b) A patient with a C-shaped PTV with lungs and spinal cord defined as OARs. (c) A patient with a large C-shaped PTV with a kidney and cauda equina defined as OARs.

3.3 Results

For all of the PARETO runs presented in this study, five coplanar beams were defined and a population of 300 was optimized to 1000 generations using 12 CPUs unless otherwise stated. In some cases, we present multiple runs to show the reproducibility of these results. However, it is not practical to test each test for reproducibility since run times can vary between approximately 15 – 50 hours (see Fig. 3.47).

3.3.1 Demonstration of the effect of an optimized BEV margin parameter on the spine phantom

For each beam, our standard approach is to define a margin parameter that controls the size of the kernel used to dilate or erode the projection of the PTV in the BEV (see Section 3.2.2.7). To show the usefulness of this approach, we compare a run done without modifying the PTV projection in the BEV to a run done using the optimized margin parameter on the spine phantom. A single linear gradient was used for PTV modulation (Equations 3.4 and 3.5) and OAR reduction parameters were applied in each case (Section 3.2.2.6). Figure 3.8 shows the tradeoffs of PTV fitness versus left kidney and spinal cord fitness (Fig. 3.8a, b). For both projections, the run done with the optimized BEV margin parameter achieves solutions of superior PTV conformity (low PTV fitness). A merged optimal set was constructed from both runs following the methodology of Section 3.2.3 (Fig. 3.8c, d). The run done with the optimized margin parameter has a survival fraction (see Equation 3.27) of solutions in the merged set of 96% compared to only 4% for the run done without (Fig. 3.8e).

Also, to show the advantage of modifying the beam width on concave geometries, we present the beam intersection volume (BIV) of 20 beams oriented tangential to the concavity of the spine phantom PTV (see Fig. 3.9). When beams of constant width are used, points closer to the concavity have a lower BIV (Fig. 3.9a). However, when Ferret is used to optimize the beam widths, the highest BIV is obtained near the centre of the PTV, and the standard deviation of voxels in the PTV decreases from 3.01 to 1.95, an improvement of 35% (Fig. 3.9b). For these results, beam widths were defined as parameters bound within the range $[0.5R_{conc}, 1.5R_{conc}]$, where R_{conc} is the difference between the outer and inner radii of the concave PTV. Also, beam orientations were held constant, all beams were weighted equally, and the objective function was defined as the standard deviation of the BIV within the PTV.

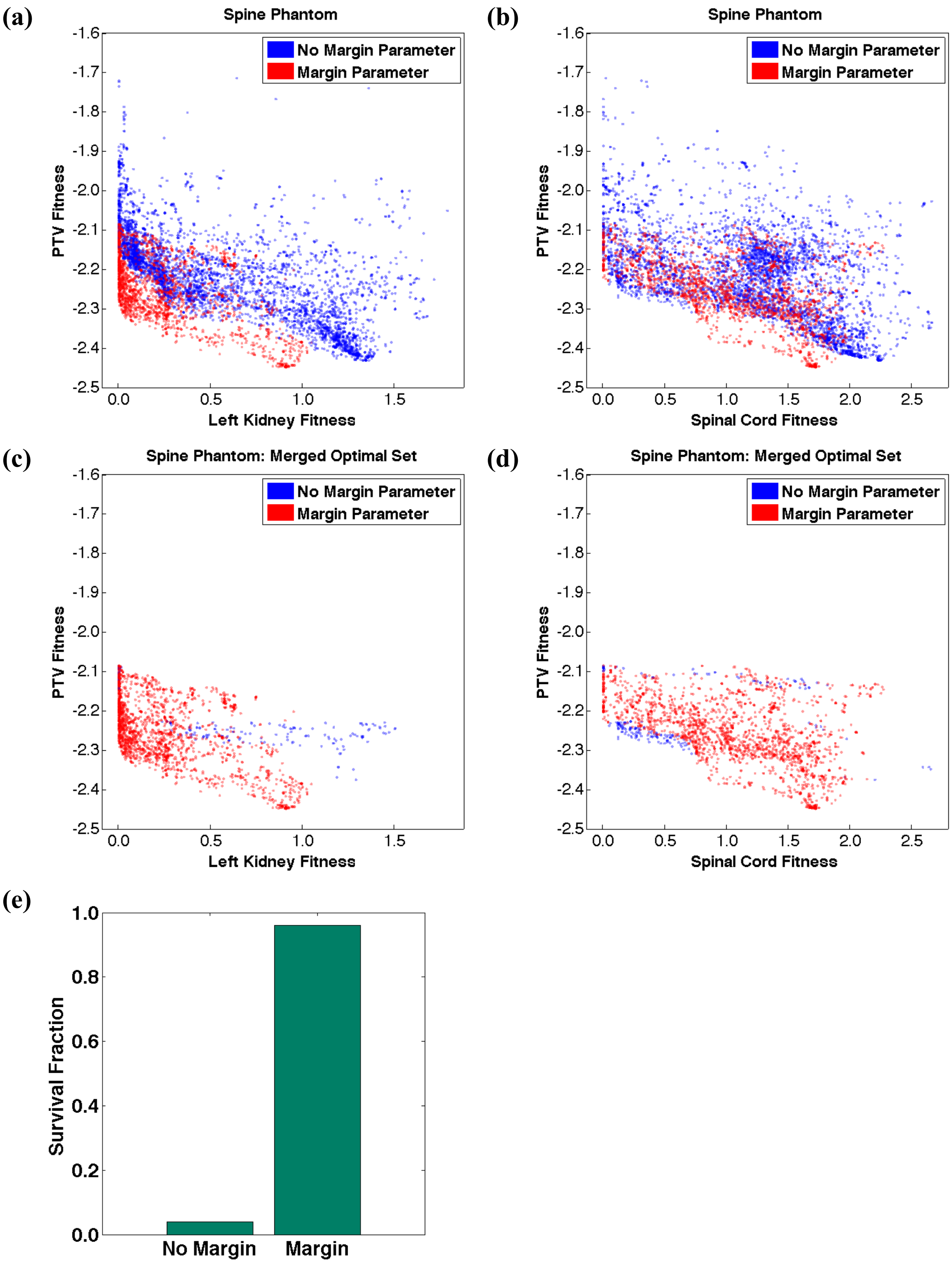

Figure 3.8: (a, b) The set of non-dominated solutions from a 5 beam run done with a single linear gradient for PTV modulation on the spine phantom with no optimized BEV margin parameter (blue) and with the optimized BEV margin parameter (red). Tradeoff surfaces are shown for PTV conformity fitness versus OAR dose. (c, d) The set of non-dominated solutions constructed by applying Ferret's tournament selection operator to the combination of all solutions from each run. Each solution that survives the tournament is colour-coded according to the run that generated it. (e) The survival fraction of each run in the merged set.

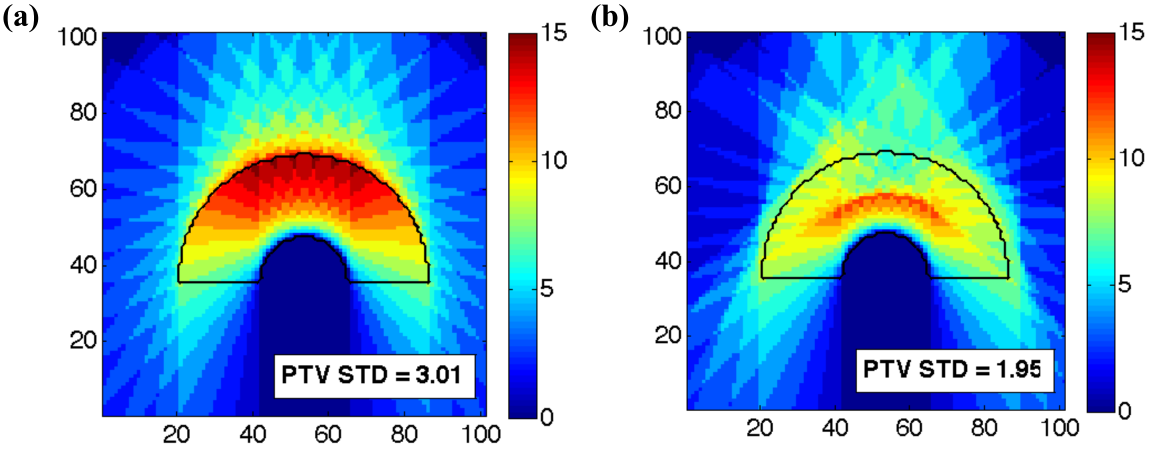

Figure 3.9: 20 beams are defined at set beam angles tangential to the concavity of the PTV. Each beam has the same flat fluence distribution (a constant value of one). These are *not* solutions from PARETO runs, and no attenuation or patient scatter kernel is used here. (a) The BIV obtained when the beam width is held constant. (b) Ferret is used to optimize each beam width so as to minimize the STD of the BIV.

3.3.2 Comparison of the coupled and decoupled parameterizations of the linear gradient method

Equation 3.4 and 3.6 demonstrate two different ways to specify a linear gradient function across a 2D map. We have tested both the coupled and the decoupled parameterizations with a single linear gradient function for PTV modulation. Standard OAR reduction parameters were applied in each case (see Section 3.2.2.6). Figures 3.10, 3.11 and 3.12 show the results for the spine phantom, lung patient, and cauda equina patient, respectively. The coupled and the decoupled runs show similar projections in the tradeoffs of PTV fitness versus OAR dose for all patient geometries.

For each patient geometry, two independent merged optimal sets were constructed using Ferret's tournament selection operator by pairing a run done with the coupled and a run done with the decoupled parameterization. The tradeoff projections for one merged set of each geometry are shown in Figures 3.10, 3.11 and 3.12, where each solution is colour-coded according to the particular parameterization that generated it. For the lung patient, the coupled method produces superior well-balanced solutions (those that do well in both PTV conformity and OAR dose) (Fig. 3.11c, d). The decoupled method, on the other hand, produces more non-dominated solutions near the extrema of low OAR dose or low PTV fitness. For the cauda equina patient, the opposite trend is visible on the tradeoff of PTV fitness versus right kidney fitness: the decoupled method produces superior well-balanced solutions and the coupled method produces superior solutions near the extrema (Fig. 3.12c). However, in the projection of PTV fitness versus cauda equina fitness, the coupled method shows both well-balanced and extreme non-dominated solutions, while the decoupled method shows non-dominated solutions only at the extreme of low PTV fitness (Fig. 3.12d). There are no obvious trends visible for the spine phantom (Fig. 3.10c, d).

Figures 3.10e, 3.11e and 3.12e show the survival fraction (Equation 3.27) of solutions in the two merged optimal sets formed for each patient geometry. We note that both the coupled and the decoupled method have a higher survival fraction in at least one merged set for both the lung patient and cauda equina patient. Thus, there is no conclusive trend. For the spine phantom, the decoupled method has a higher survival fraction in both sets (Fig. 3.10e).

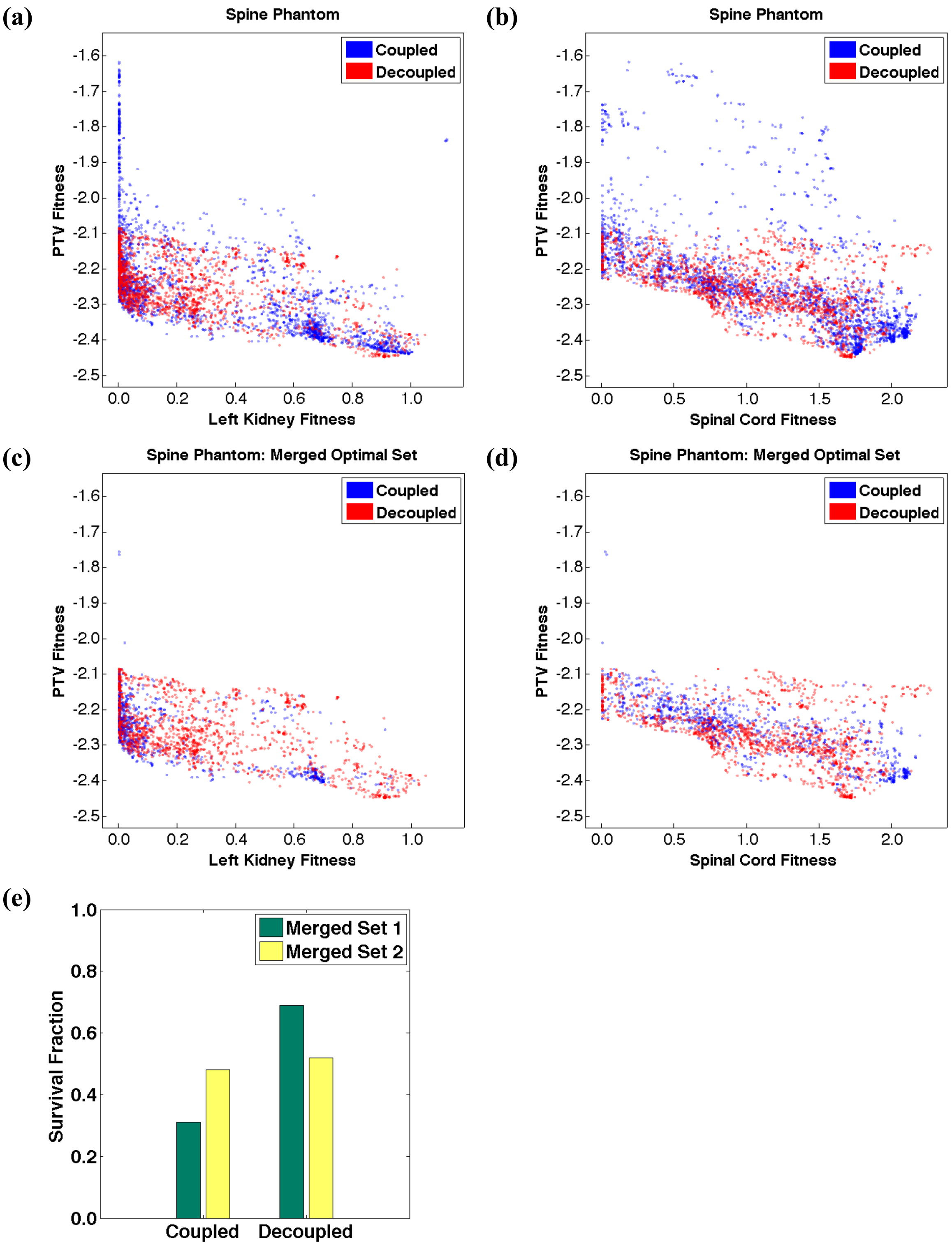

Figure 3.10: (a, b) Sets of non-dominated solutions from 5 beam runs done with the coupled (blue) and decoupled (red) parameterizations of the single linear gradient method for the spine phantom. Tradeoff surfaces are shown for PTV conformity fitness versus OAR dose. (c, d) The set of non-dominated solutions constructed by applying Ferret's tournament selection operator to the combination of all solutions from the coupled and decoupled runs. Each solution that survives the tournament is colour-coded according to the parameterization that was used to generate it. (e) The survival fraction of each run. Two merged sets are created from two runs done with each parameterization (the tradeoffs of the first set are shown in a and b).

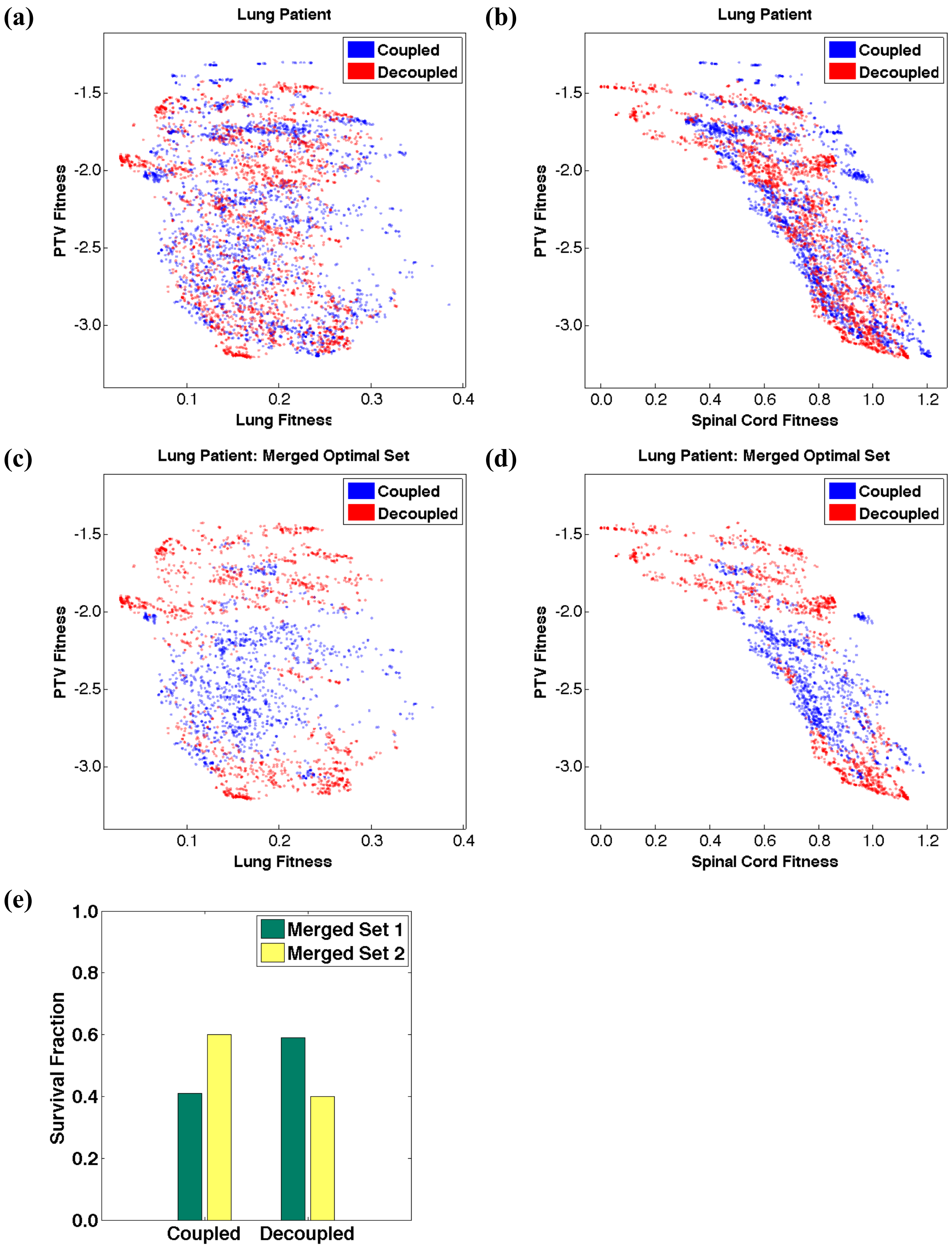

Figure 3.11: (a, b) Sets of non-dominated solutions from 5 beam runs done with the coupled (blue) and decoupled (red) parameterizations of the single linear gradient method for the lung patient. Tradeoff surfaces are shown for PTV conformity fitness versus OAR dose. (c, d) The set of non-dominated solutions constructed by applying Ferret's tournament selection operator to the combination of all solutions from the coupled and decoupled runs. Each solution that survives the tournament is colour-coded according to the parameterization that was used to generate it. (e) The survival fraction of each run. Two merged sets are created from two runs done with each parameterization (the tradeoffs of the first set are shown in a and b).

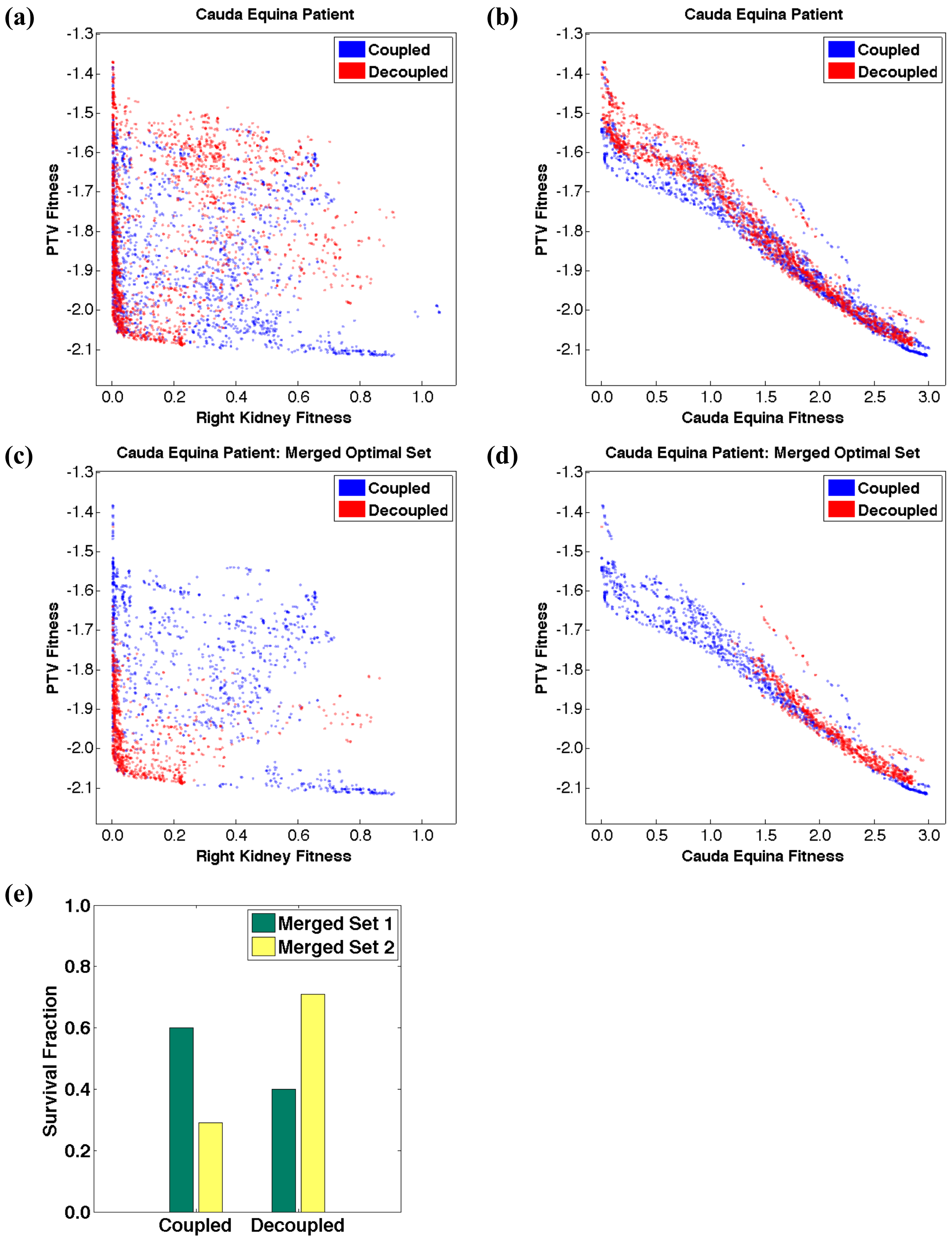

Figure 3.12: (a, b) Sets of non-dominated solutions from 5 beam runs done with the coupled (blue) and decoupled (red) parameterizations of the single linear gradient method for the cauda equina patient. Tradeoff surfaces are shown for PTV conformity fitness versus OAR dose. (c, d) The set of non-dominated solutions constructed by applying Ferret's tournament selection operator to the combination of all solutions from the coupled and decoupled runs. Each solution that survives the tournament is colour-coded according to the parameterization that was used to generate it. (e) The survival fraction of each run. Two merged sets are created from two runs done with each parameterization (the tradeoffs of the first set are shown in a and b).

3.3.3 Investigation into the effect of varying the number of PTV fluence parameters on fluence complexity and solution quality

For all of the parameterizations described in Section 3.2.2, one would expect that increasing the number of parameters should increase the complexity of the fluence modulation. Nonetheless, this assumption must be tested. Not all parameterizations are expected to exhibit the same relationship between fluence complexity and the number of parameters (some may show a near-linear upward trend, while others may not). Furthermore, some methods may produce fluence maps that are overall more modulated compared to those produced by other methods. Therefore, when a high degree of fluence complexity is desirable for good solution quality on a certain patient geometry, it may be more difficult for the GA to search the space for certain parameterizations, even if a large number of parameters is used. We have tested the relationship between fluence complexity and number of parameters for each method using randomly generated fluence maps, as well as by performing PARETO runs for several patient geometries. For the PARETO runs, standard OAR reduction parameters were used unless otherwise stated (Section 3.2.2.6). Tools for analyzing fluence complexity and solution quality are discussed in Section 3.2.3.

3.3.3.1 Linear Gradient Method

Using the decoupled parameterization of a linear gradient function (Equations 3.4 and 3.5), we generated 5000 fluence maps using random parameter values for each of one to four multiplied gradients. Figure 3.13 shows how the modulation index and gradient entropy vary with the number of multiplied linear gradients. We note that the mean modulation index and gradient entropy only increase up to two gradients and decrease after that.

We performed several PARETO runs with one to four multiplied gradients for PTV modulation. Figures 3.14−3.17 show the results for the spine phantom, lung patient, and cauda equina patient. For the spine phantom, the non-dominated solutions of each run overlap in the projections of PTV versus OAR fitness (Fig. 3.14a, b). However, the run done with a single linear gradient achieves superior solutions with respect to PTV fitness, and achieves the highest mean survival fraction in the round robin tournament (Fig. 3.14c). The round robin tournament also shows that there is a decrease in solution quality with an increase in the number of gradients. Nonetheless, we note that the mean modulation index and gradient entropy show a modest increase with the number of gradients used (Fig. 3.14d, f). The mean fractal dimension, however, shows no clear trend (Fig. 3.14e).

Figure 3.15 shows the result of varying the number of multiplied gradients used for fluence modulation on the lung patient. The tradeoff of PTV versus lung fitness shows that the run done with two gradients achieves superior solutions in lung dose (Fig. 3.15a). Figure 3.15b shows that runs done with two or three gradients achieve superior well-balanced solutions in PTV fitness and spinal cord dose. The results of round robin tournament support the visual observations that the two and three gradient runs are superior (they come in first and second place, respectively) (Fig. 3.15c). Also, the fluence metrics show that the mean modulation index increases with the number of gradients up to three gradients (Fig. 3.15d), while the mean gradient entropy appears to saturate at only two gradients (Fig. 3.15f). The mean fractal dimension stays roughly constant (Fig. 3.15e).

As a test of the reproducibility of these results, we performed five sets of runs on the lung patient with one to four multiplied gradients in each set. Five round robin tournaments were thus generated, and the results are shown in Figure 3.16. We note that although it is difficult to see a trend that is consistent across each set of runs and the highest quality run varies considerably for each tournament, runs with two gradients win the round robin tournament in two cases.

Finally, Figure 3.17 indicates that using two multiplied gradients also improves solution quality on the cauda equina patient. The tradeoff of PTV versus right kidney fitness shows that the solutions of the two gradient run are overall lower in right kidney dose, while some solutions simultaneously achieve superior PTV conformity as well (Fig. 3.17a). The tradeoff of PTV versus cauda equina fitness shows that the solutions of the two gradient run are overall lower in PTV fitness while comparable in cauda equina dose (Fig. 3.17b). Also, the round robin tournament confirms that the two gradient run is superior (Fig. 3.17c). For this patient geometry, the mean modulation index and gradient entropy of the two and four gradient runs are highest (Fig. 3.17d, f). The mean fractal dimension again stays roughly constant with the number of gradients (Fig. 3.17e).

However, we note that when the modulation index and gradient entropy are calculated prior to the application of OAR reduction parameters, there is a strong upward trend in the gradient entropy with an increasing number of gradients, and a smaller trend in the modulation index up to three gradients (Fig. 3.18).

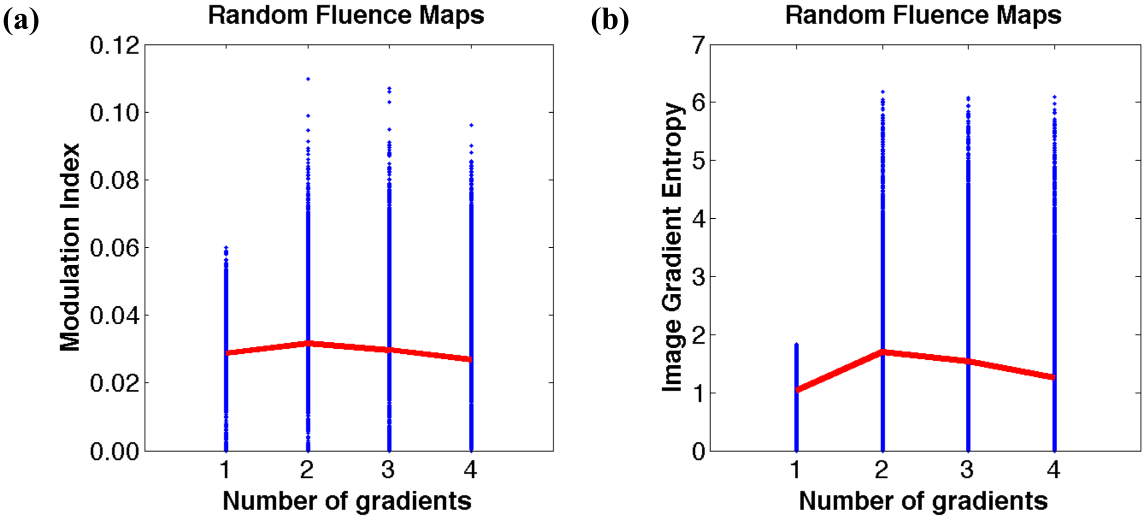

Figure 3.13: The modulation index and image gradient entropy of 5000 random fluence maps (60x60 pixels) generated with a varying number of multiplied linear gradients (blue). The red curve joins the mean values.

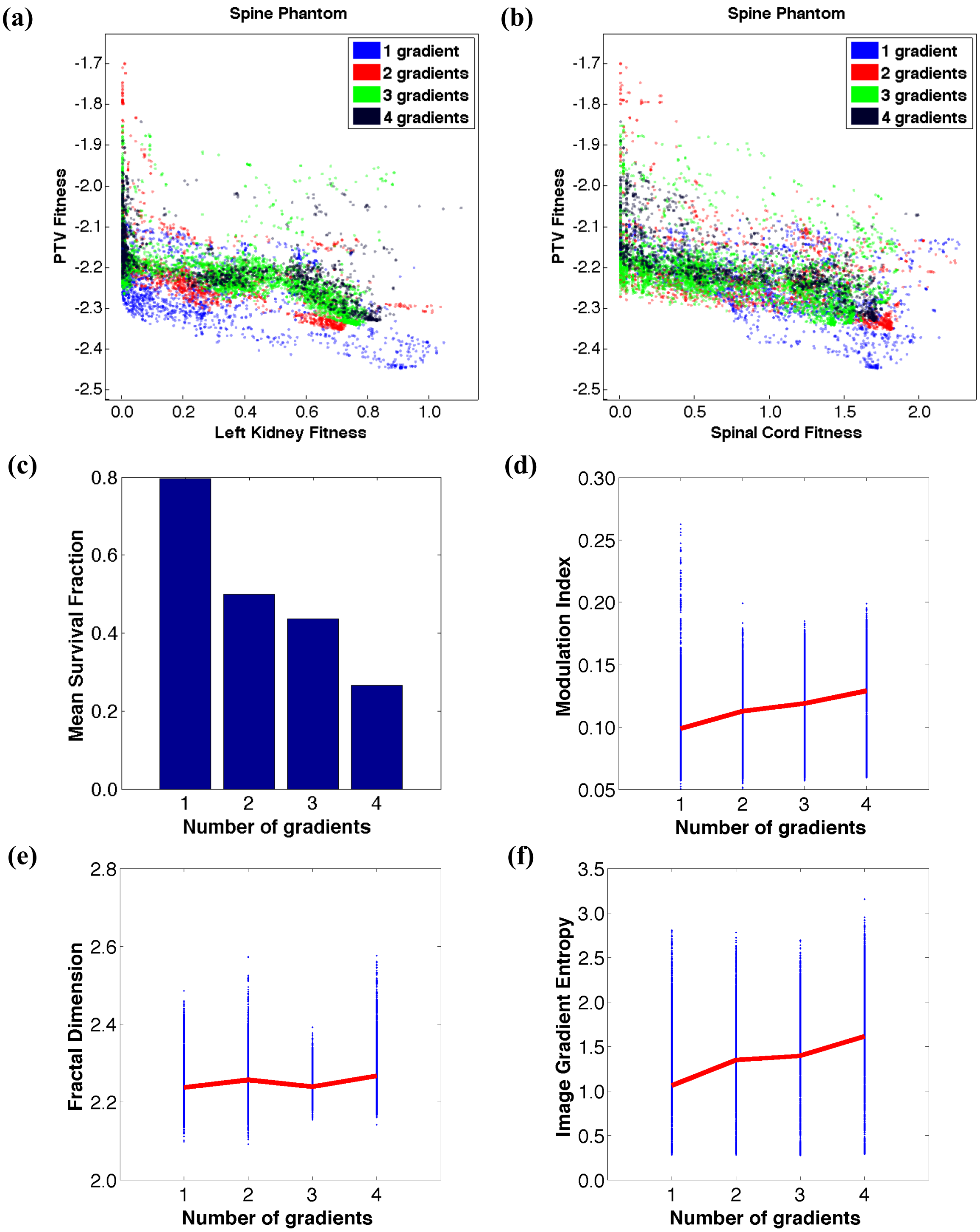

Figure 3.14: (a, b) Optimal solutions for the spine phantom from 5 beam runs done with a varying number of gradients as specified by the decoupled parameterization of the linear gradient method. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

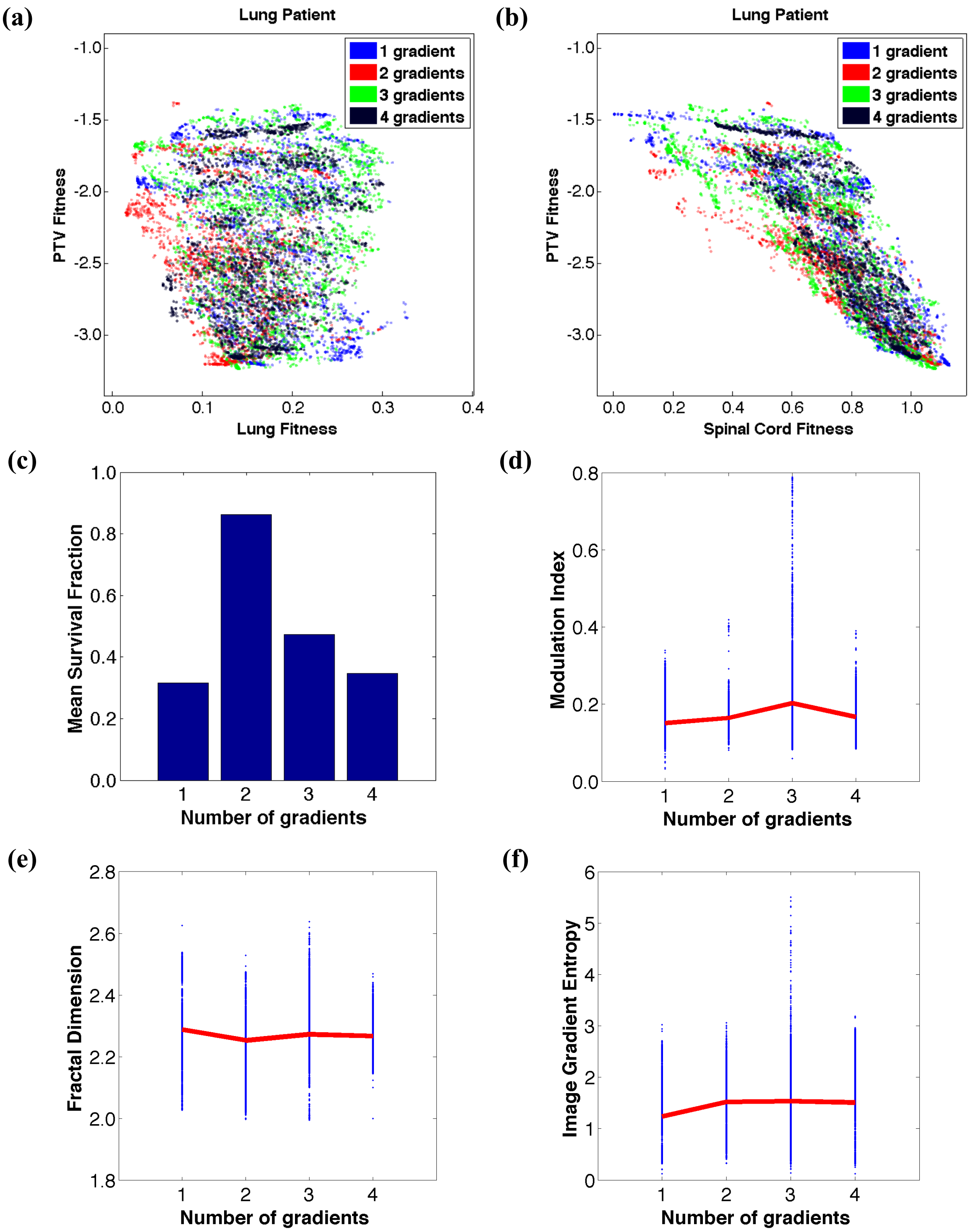

Figure 3.15: (a, b) Optimal solutions for the lung patient from 5 beam runs done with a varying number of gradients as specified by the decoupled parameterization of the linear gradient method. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

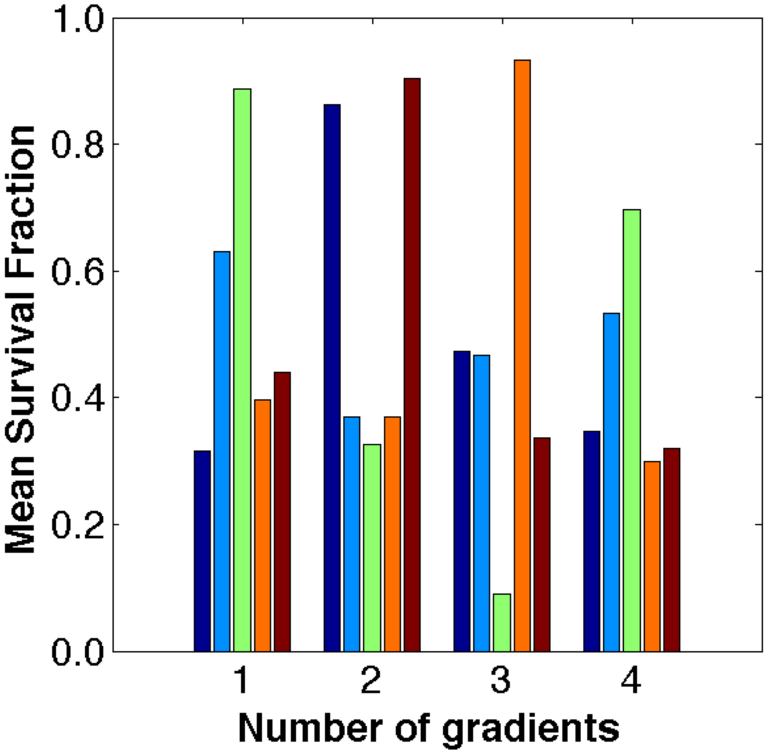

Figure 3.16: Scores from five round robin tournaments for five 5 beam runs done with a varying number of multiplied linear gradients on the lung patient. The round robin tournament shown in navy is also depicted in Figure 3.15.

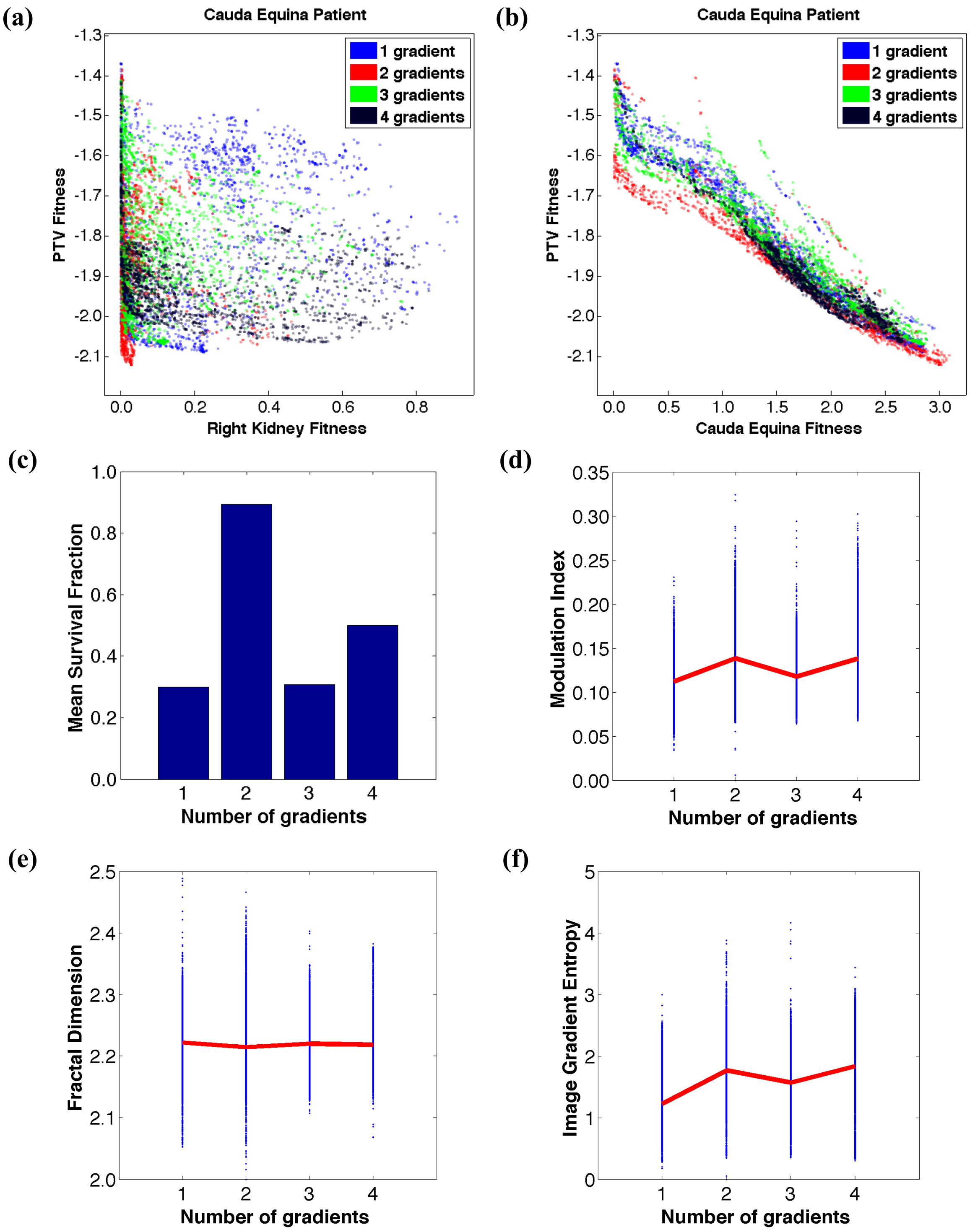

Figure 3.17: (a, b) Optimal solutions for the cauda equina patient from 5 beam runs done with a varying number of gradients as specified by the decoupled parameterization of the linear gradient method. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

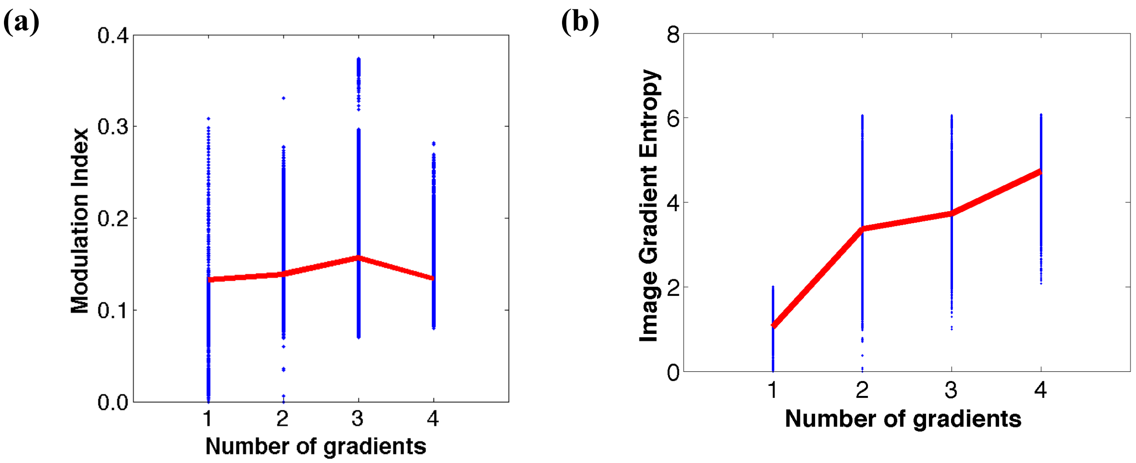

Figure 3.18: (a) Modulation index and (b) gradient entropy of the runs shown in Figure 3.17 prior to the application of OAR reduction parameters.

3.3.3.2 Cosine Transform Method

We generated 5000 random fluence maps using the cosine transform method to produce a modulated fluence pattern over the PTV projection in each BEV with frequency radii of 1, 2, 3, 4, and 8 as specified in Equation 3.14 (3, 6, 11, 17, and 58 frequencies, respectively). Figure 3.19 shows the modulation index and gradient entropy of these maps versus the number of frequencies used. The mean modulation index shows an increase with the number of frequencies, and jumps significantly at 58 frequencies. The gradient entropy is quite low for the maps with 3 frequencies, and shows no clear trend for higher numbers of frequencies.

Several PARETO runs were also performed with frequency radii of 1, 2, 3, 4, and 8. For the spine phantom, the tradeoffs of PTV versus left kidney fitness and PTV versus spinal cord fitness show that runs done with a smaller number of frequency parameters achieve superior PTV conformity (Fig. 3.20a, b). The round robin tournament also indicates a decrease in solution quality for an increase in the number of frequencies past 6 frequency parameters. The mean gradient entropy tends to increase with an increase in the number of frequency parameters (Fig. 3.20f). The mean modulation index and fractal dimension show little variation (Fig. 3.20d, e).

The results for the lung patient, on the other hand, show that solution quality increases significantly for a very high number of frequency parameters (Fig. 3.21c). The tradeoff of PTV versus lung fitness shows that the run done with 58 parameters has some success in achieving solutions that are extremely low in lung dose (Fig. 3.21a). However, it is more difficult to see the relationship between the other runs on the tradeoff plots (Fig. 3.21a, b). The spread of the fractal dimension and gradient entropy generally tends to increase with the number of frequency parameters (Fig. 3.21e, f). Also, the mean gradient entropy shows no clear trend at low numbers of frequency parameters but does show a significant increase at 58 parameters (Fig. 3.21f). The mean modulation index shows a gradual increase with the number of frequency parameters but saturates quickly (Fig. 3.21d).

The same general behaviour is seen for runs done with a varying number of frequency parameters on the cauda equina patient. However, in this case, the run done with 58 frequency parameters drops slightly in overall solution quality compared to the run done with 17 frequency parameters (Fig. 3.22c). This may be due to the fact that the solutions of the run done with 17 frequency parameters are slightly superior in PTV conformity (Fig. 3.22b). The mean gradient entropy increases up to 58 frequency parameters (Fig. 3.22f), while the modulation index saturates more quickly at 11 parameters (Fig. 3.22d). The mean fractal dimension remains approximately constant except for a noticeable increase at 58 parameters (Fig. 3.22e).

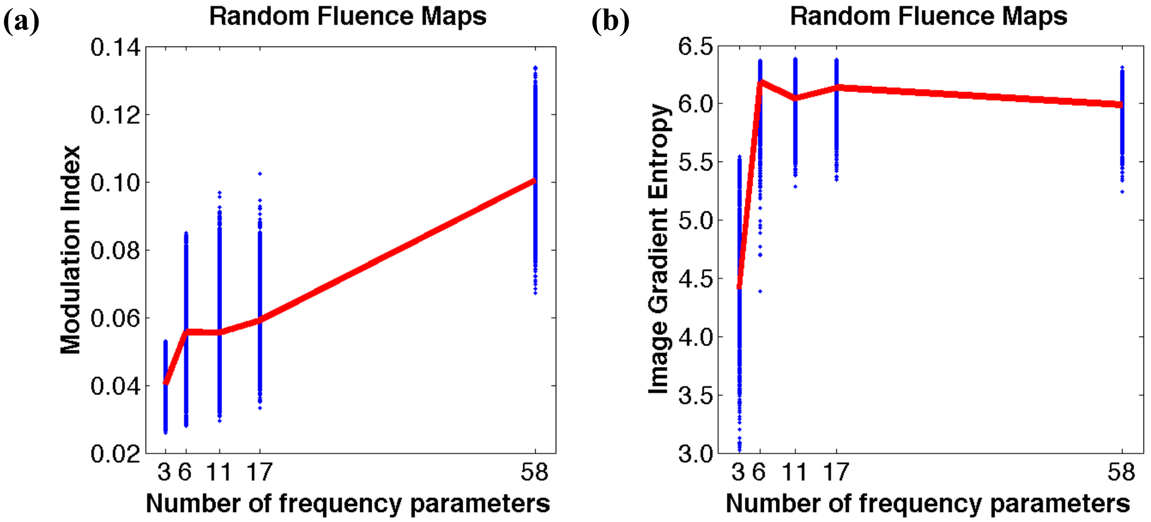

Figure 3.19: The modulation index and image gradient entropy of 5000 random fluence maps (60x60 pixels) generated with a varying number of frequency parameters (blue). The red curve joins the mean values.

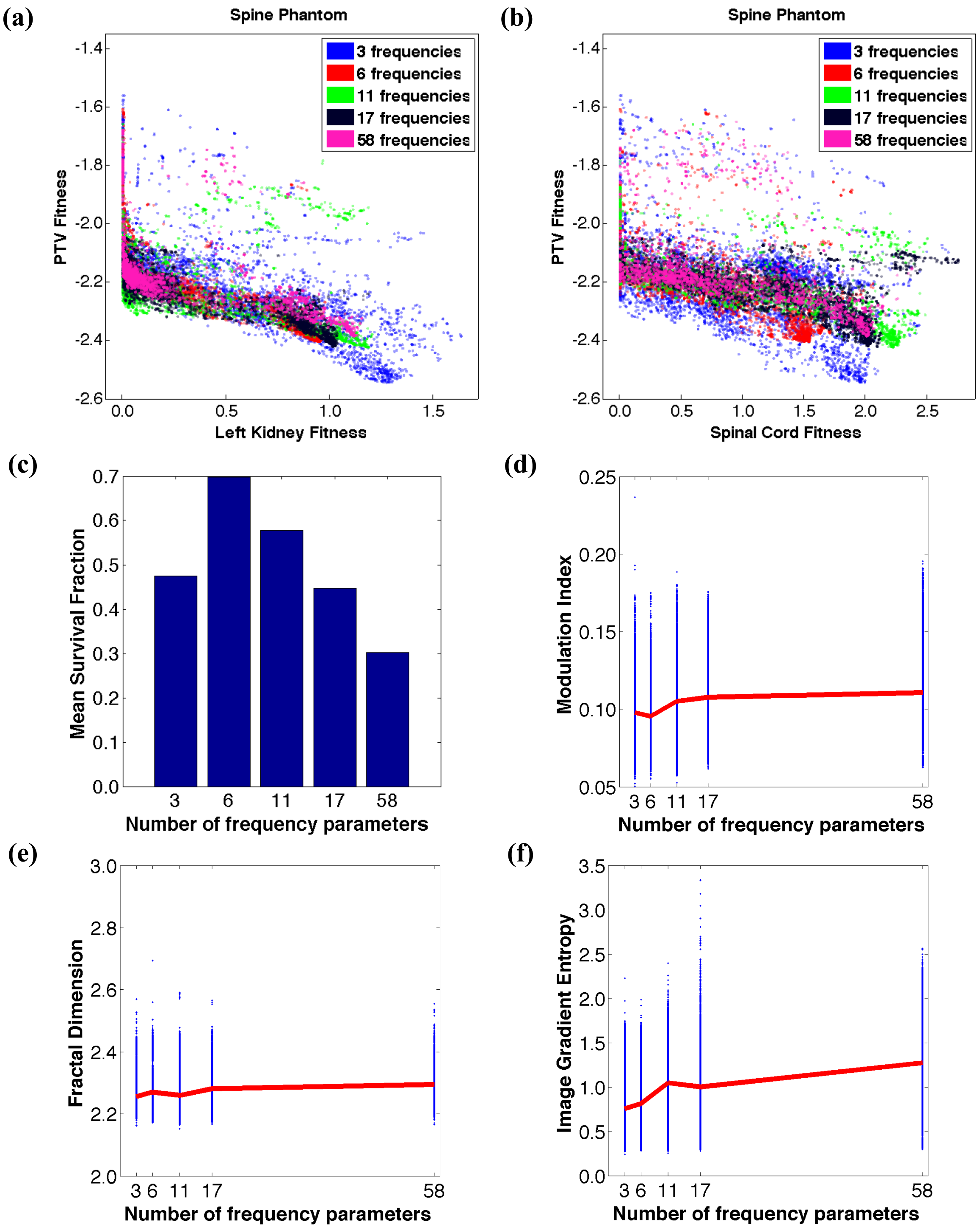

Figure 3.20: (a, b) Optimal solutions for the spine phantom from 5 beam runs done with a varying number of frequency parameters as specified by the cosine transform method. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

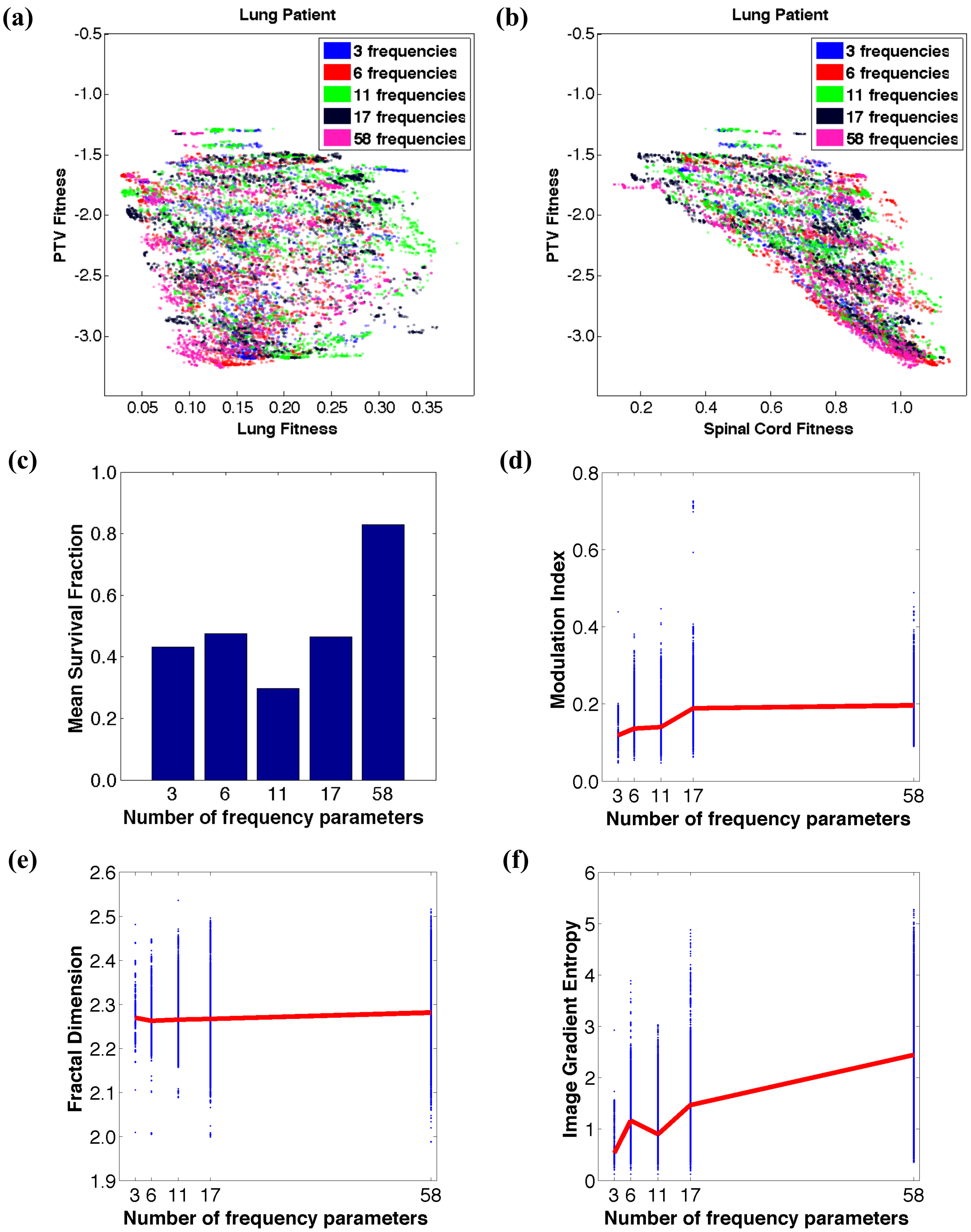

Figure 3.21: (a, b) Optimal solutions for the lung patient from 5 beam runs done with a varying number of frequency parameters as specified by the cosine transform method. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

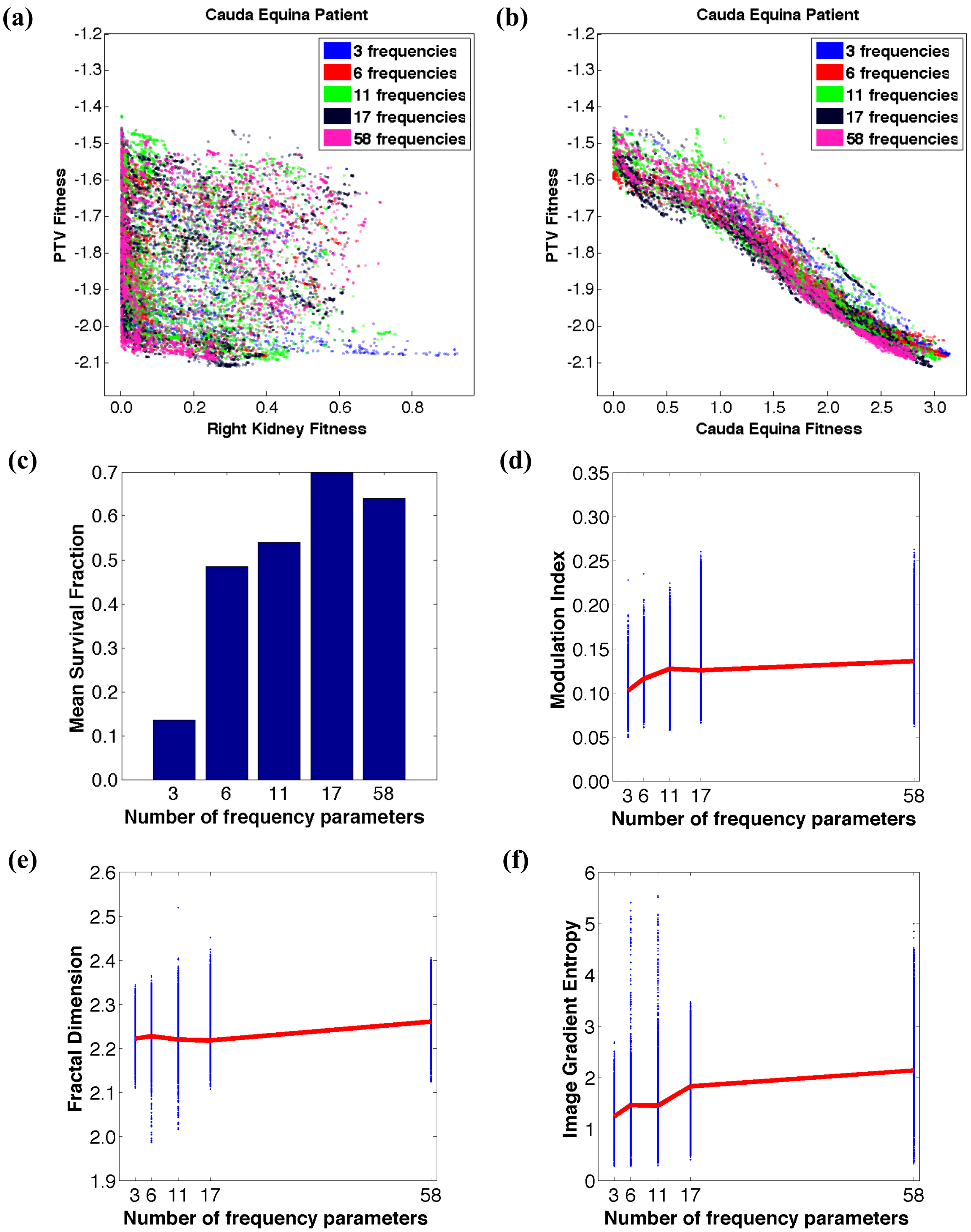

Figure 3.22: (a, b) Optimal solutions for the cauda equina patient from 5 beam runs done with a varying number of frequency parameters as specified by the cosine transform method. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

3.3.3.3 Beam Group Method

In this section, we investigate the effect of changing the resolution of the beam group grid. We first generate 5000 fluence maps for each grid size (9, 25, 49, 81 beam groups) using random parameter values. Figure 3.23 shows the modulation index and gradient entropy of the fluence maps versus the number of beam groups. The modulation index shows a clear, nearly linear trend of an increase in the mean modulation index for an increase in the number of beam groups. The mean gradient entropy subtly increases with the number of beam groups and appears to saturate, while the range of the gradient entropy tends to decrease.

In the first set of PARETO runs, OAR reduction parameters are applied (Fig. 3.24, 3.25, 3.26, 3.27, 3.28). For the spine phantom, we see a clear decrease in solution quality with an increase in the number of beam groups both by visualizing the tradeoffs of PTV versus left kidney and spinal cord fitness (Fig. 3.24a, b), and by the results of the round robin tournament (Fig. 3.24c). Runs done with a smaller number of beam groups achieve solutions of superior PTV conformity, resulting in greater success in the round robin tournament. Figure 3.24 also shows that an increase in the number of beam groups directly translates into an increase in the mean modulation index and gradient entropy of the fluence maps (Fig. 3.24d, f).

Results for the lung patient show an increase in overall solution quality with an increase in the number of beam groups until 49 beam groups (Fig. 3.25c). Indeed, the 49 beam group run achieves solutions of superior PTV conformity in the tradeoff of PTV versus spinal cord dose (Fig. 3.25b). However, solution quality decreases for 81 beam groups, which may be explained by the observation that this run does worse than all others in producing solutions that are both very well conformed to the PTV and extremely low in lung dose on the tradeoff of PTV versus lung fitness (Fig. 3.25a). The mean modulation index, fractal dimension, and gradient entropy all increase steadily with an increase in the number of beam groups, though the gradient entropy appears to saturate around 49 beam groups (Fig. 3.25d, e, f).

Figure 3.26 shows the results from five independent round robin tournaments for runs done with a varying number of beam groups on the lung patient. In two cases, 49 beam groups win a round robin tournament, while in three cases, 25 beam groups win. Furthermore, for all tournaments, the mean survival fraction increases up to 25 or 49 beam groups, and then decreases again at 81 beam groups. Therefore, Figure 3.26 shows that trends in the round robin scores of runs done with a varying number of beam groups are fairly reproducible, and thus a moderate to high number of beam groups reliably produces the highest solution quality.

For the cauda equina patient, a run done with 25 beam groups produces solutions of superior PTV conformity than is achieved by the solutions of other runs (Fig. 3.28b). The round robin tournament confirms this success (Fig. 3.28c). The mean modulation index and gradient entropy again show a clear upward trend with the number of beam groups (Fig. 3.28d, f). The mean fractal dimension, on the other hand, stays roughly constant until 81 beam groups (Fig. 3.28e).

In an effort to separate the effects of beam group fluence modulation from the effects of OAR reduction parameters, we performed several runs where OAR reduction parameters were omitted and no other modulation method was specifically applied to the projection of the OARs in the BEV. We postulated that a clearer trend of an increase in overall solution quality with an increase in the number of beam groups would be visible on all patient geometries compared to runs done with OAR reduction parameters. In all cases, we note that the modulation index and gradient entropy of the fluence maps (Fig. 3.29d, f, 3.30d, f, 3.31d, f) follow the same general trend as for the runs done with OAR reduction parameters applied (Fig. 3.24, 3.25, 3.28d, f). However, with these runs there is a clearer increase in the mean fractal dimension with an increase in the number of beam groups (Fig. 3.29e, 3.30e, 3.31e).

For the spine phantom, Figure 3.29a shows that runs done with 9 and 25 beam groups have a larger “halo” (a broad spread of solutions that are high in PTV and left kidney fitness) compared to runs done with 49 or 81 beam groups. However, Figure 3.29b shows that the run done with 25 beam groups contains superior well-balanced solutions in the region that is simultaneously low in PTV and spinal cord fitness. Meanwhile, the solutions of the 9 beam group run are seen to be overall higher in spinal cord dose. The 49 and 81 beam group runs appear to be nearly comparable in the tradeoff projections, although solutions from the 49 beam group run are overall slightly lower in PTV fitness in the tradeoff of PTV versus spinal cord fitness (Fig. 3.29b). Thus, the results of the round robin tournament are quite reasonable, showing a significant advantage with 25 or 49 beam groups over 9 or 81 beam groups (Fig. 3.29c).

For the lung patient, a drastic increase in overall solution quality is seen for 25 beam groups compared to 9 beam groups (Fig. 3.30c). In the tradeoff of PTV versus spinal cord fitness, the solutions of the 9 beam group run are overall worse in PTV conformity compared to the other runs, while the 25 beam group run contains very well conformed solutions (Fig. 3.30b). The 49 and 81 beam group runs are comparable and also inferior to the 25 beam group run (Fig. 3.30c).

The tradeoff surfaces for the cauda equina patient reveal a clear advantage in solution quality with 25, 49, or 81 beam groups compared to only 9 beam groups. The 9 beam group run has a larger "halo" in the projection of PTV versus right kidney fitness (Fig. 3.31a) and the solutions are much higher in PTV fitness compared to all other runs in the projection of PTV versus cauda equina fitness (Fig. 3.31b). Thus, the 9 beam group run loses the round robin tournament, while the other runs are perform similarly (Fig. 3.31c).

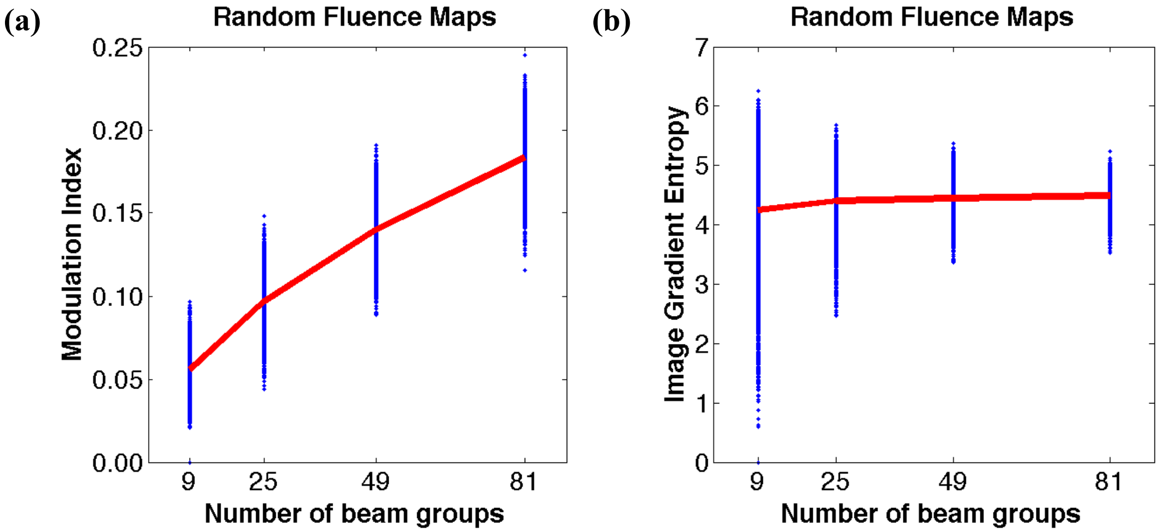

Figure 3.23: The modulation index and image gradient entropy of 5000 random fluence maps (60x60 pixels) generated with a varying number of beam groups (blue). The red curve joins the mean values.

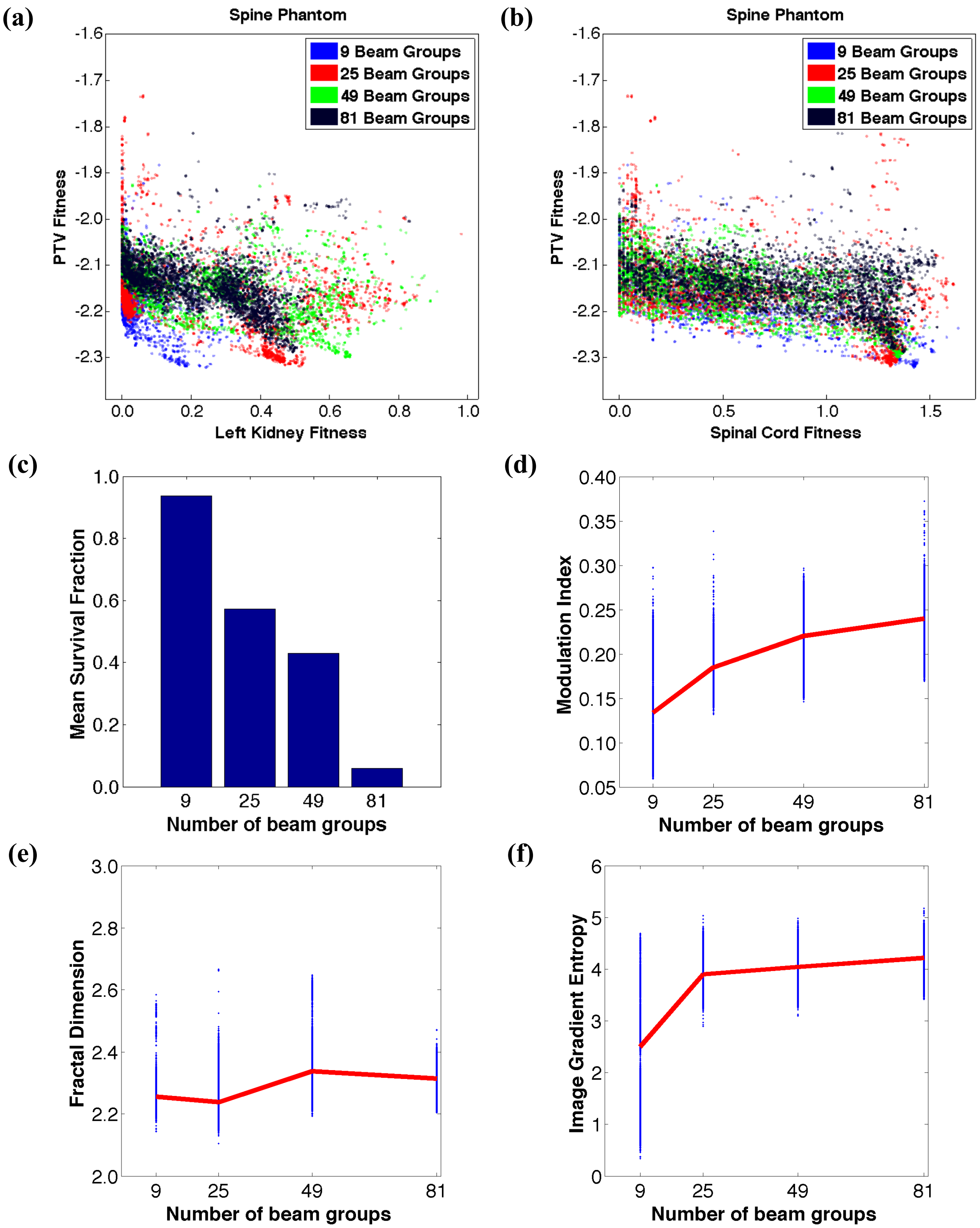

Figure 3.24: (a, b) Optimal solutions for the spine phantom from 5 beam runs done with a varying number of beam groups. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

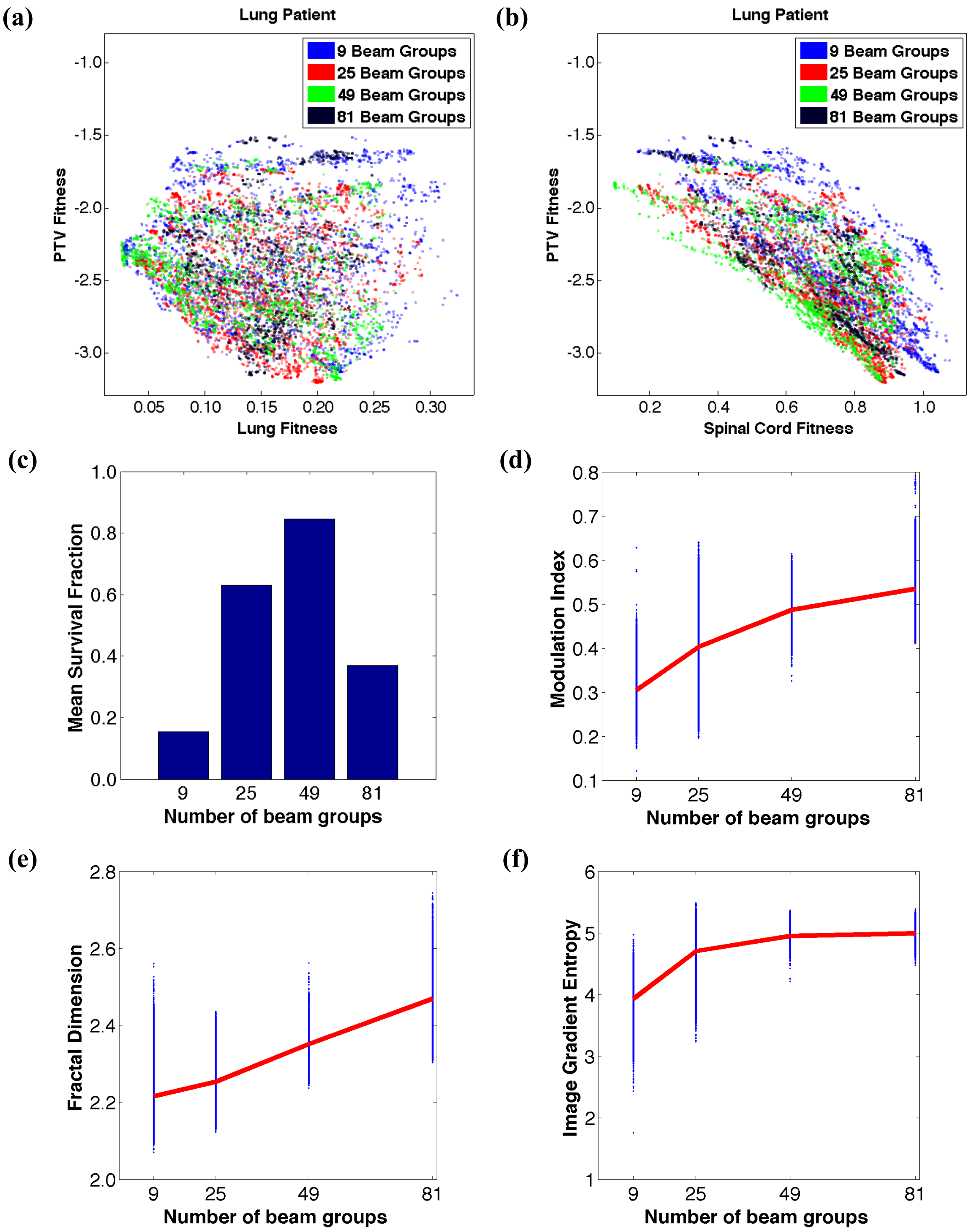

Figure 3.25: (a, b) Optimal solutions for the lung patient from 5 beam runs done with a varying number of beam groups. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

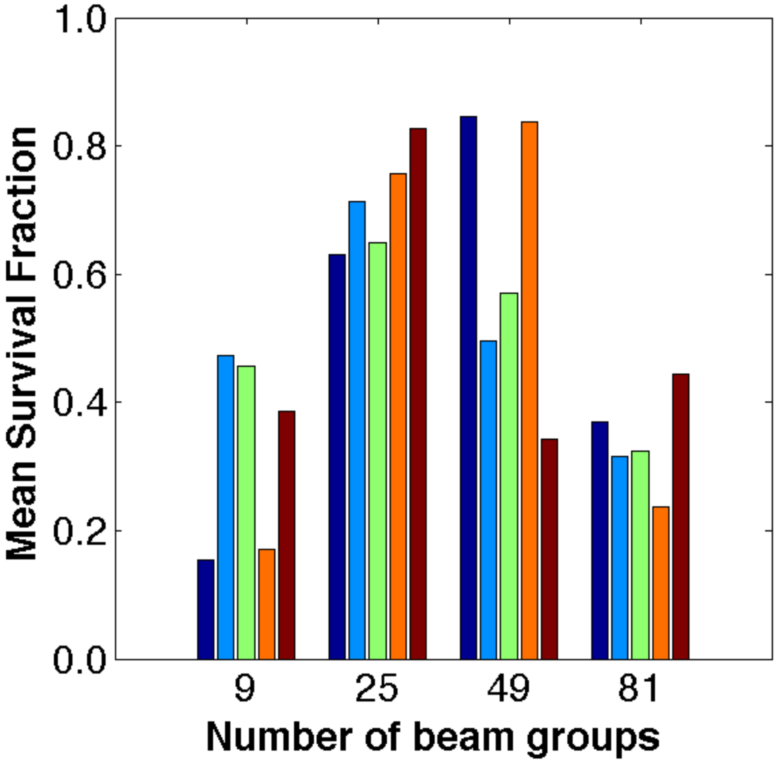

Figure 3.26: Scores from five round robin tournaments for five 5 beam runs done with each number of beam groups on the lung patient. The round robin tournament shown in navy is also depicted in Figure 3.25.

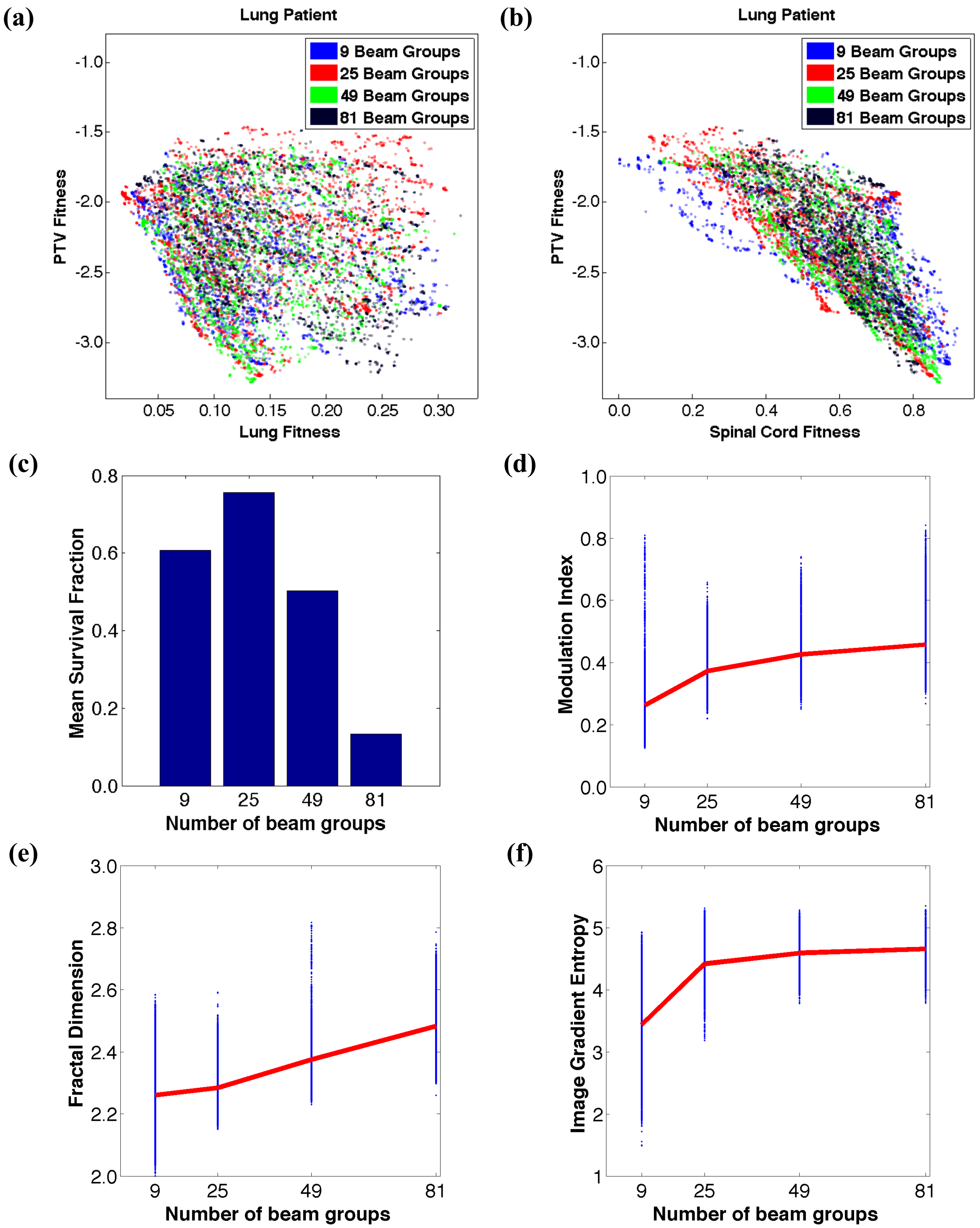

Figure 3.27: (a, b) Optimal solutions for the lung patient from 11 beam runs done with a varying number of beam groups. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

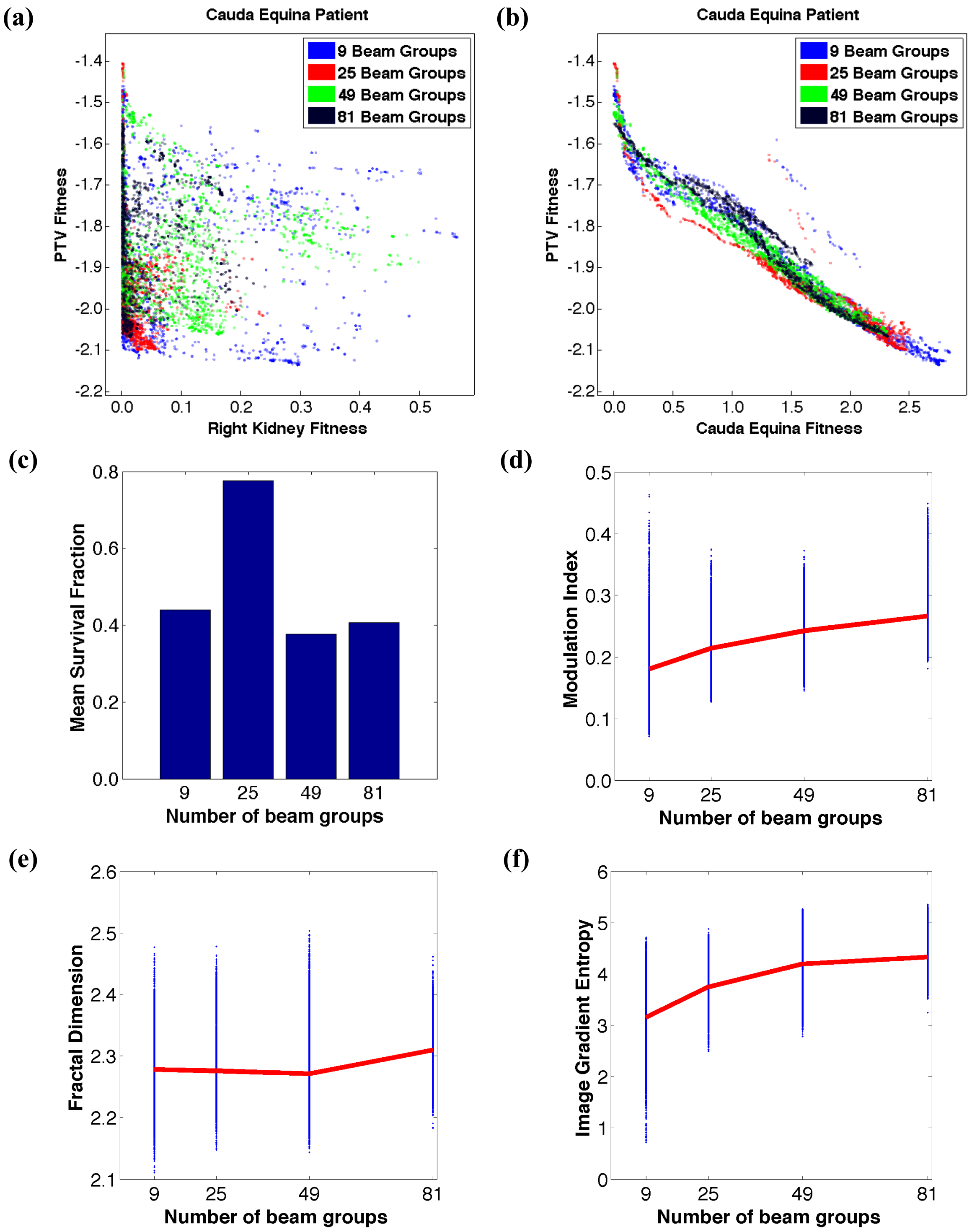

Figure 3.28: (a, b) Optimal solutions for the cauda equina patient from 5 beam runs done with a varying number of beam groups. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

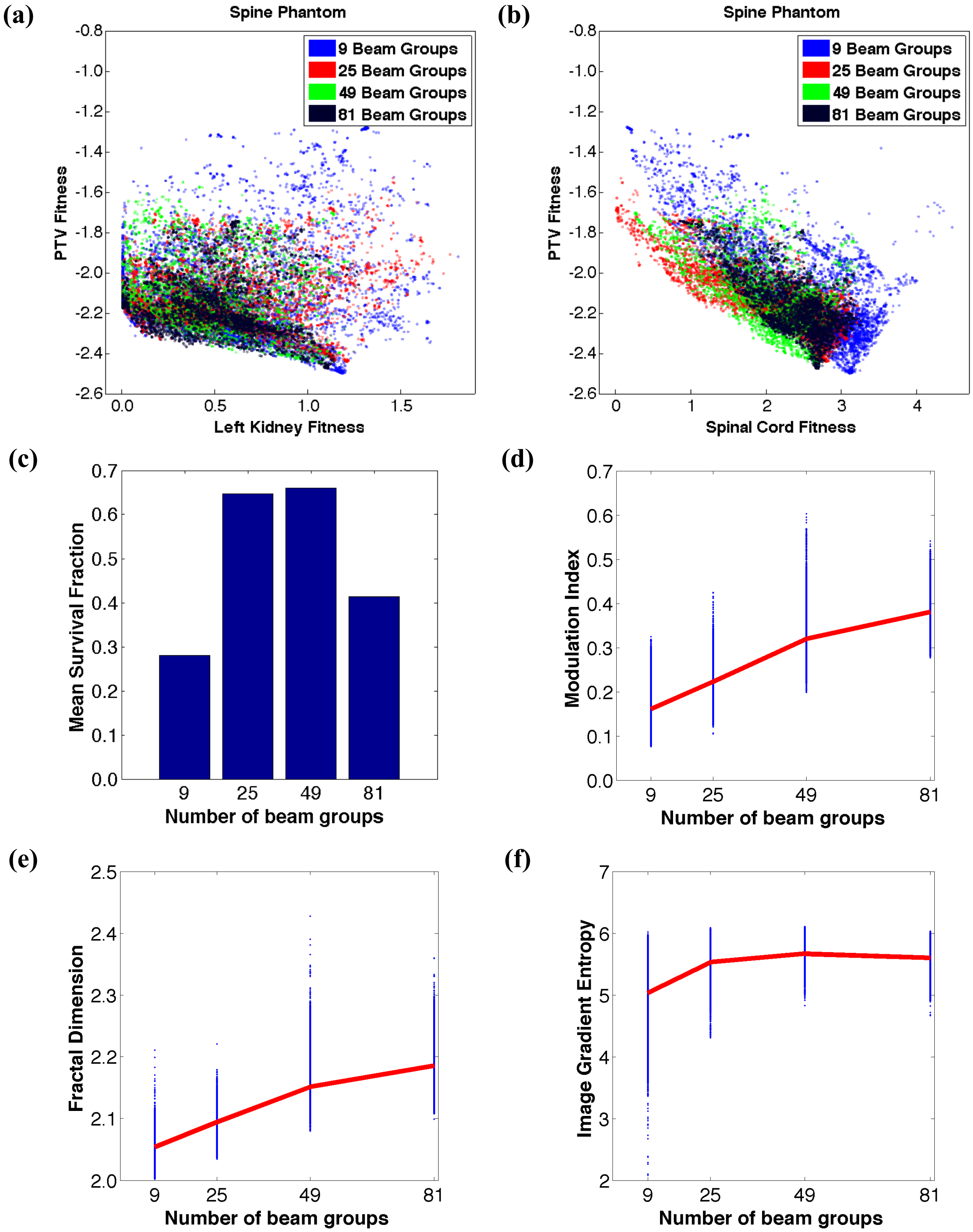

Figure 3.29: (a, b) Optimal solutions for the spine phantom from 5 beam runs done with a varying number of beam groups and no OAR reduction parameters. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

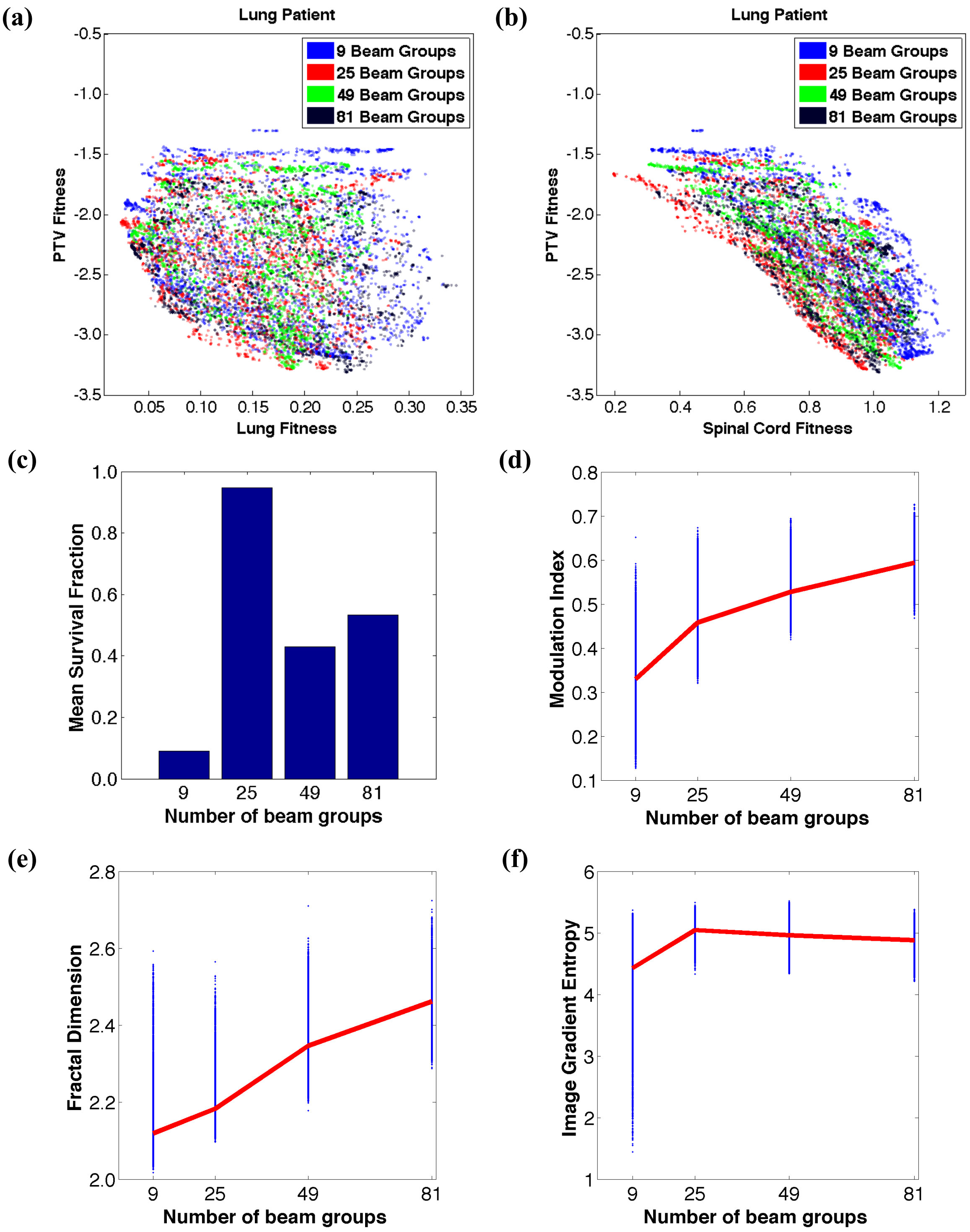

Figure 3.30: (a, b) Optimal solutions for the lung patient from 5 beam runs done with a varying number of beam groups and no OAR reduction parameters. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

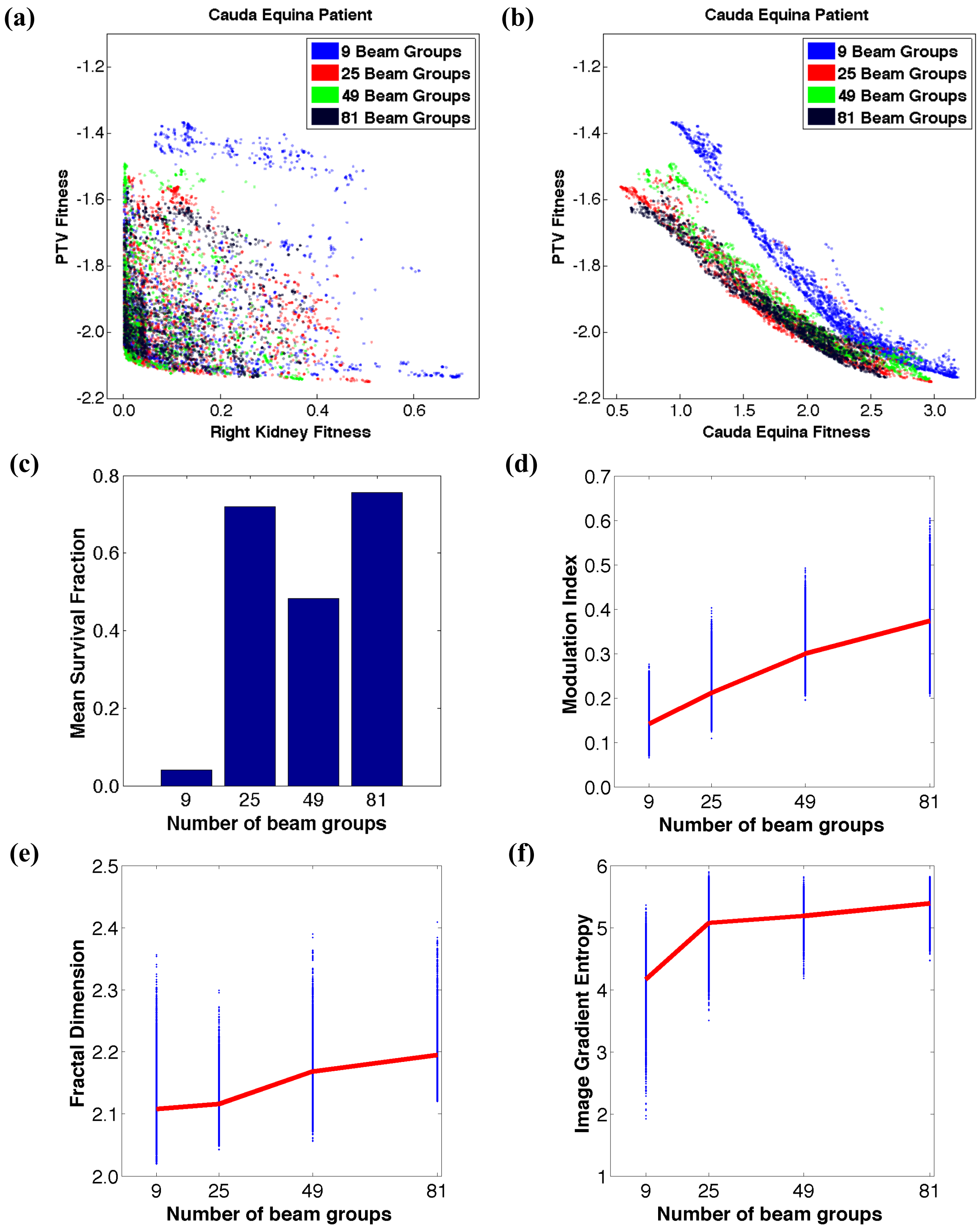

Figure 3.31: (a, b) Optimal solutions for the cauda equina patient from 5 beam runs done with a varying number of beam groups and no OAR reduction parameters. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

3.3.3.4 Isodose-based Projection Method

For the isodose-based contour method, changing the number of isodose surfaces calculated affects the number of regions that are defined for modulation on the projection of the PTV in the BEV. To show that the amount of fluence modulation increases for an increase in the number of parameters, we generated 5000 random solutions (beam angles and fluence parameters included) with 2, 4, 6 and 8 isodose surfaces defined on the spine phantom. Figure 3.32 shows the modulation index and gradient entropy of the fluence maps from these solutions. Both the mean modulation index and gradient entropy steadily increase with the number of parameters.

We performed several PARETO runs with 2, 4, 6 and 8 isodose surfaces. For the spine phantom, as little as two isodose surfaces are needed to produce superior solution quality, though runs done with higher numbers of isodose projections are not far behind (Fig. 3.33c). The modulation index, fractal dimension, and gradient entropy show no clear trend with the number of isodose projections (Fig. 3.33d, e, f).

By contrast, an increase in the number of isodose projections causes an increase in overall solution quality for the lung patient (Fig. 3.34c). Figure 3.34b shows that the 6 and 8 isodose surface runs produce solutions that are slightly superior in PTV conformity. The mean modulation index, fractal dimension, and gradient entropy all increase as the number of isodose projections increases until at least 6 projections (Fig. 3.34d, e, f).

For the cauda equina patient, the spread of solutions in the projection of PTV versus right kidney fitness is broadest with the 2 isodose surface run (Fig. 3.35a). The 8 isodose surface run, on the other hand, is perhaps the least spread in right kidney dose. The tradeoff of PTV versus cauda equina fitness shows very little difference between the runs, except that the 8 isodose surface run does not appear to contain as many outliers that are high in PTV fitness as the other runs (Fig. 3.35b). Therefore, the 4, 6, and 8 isodose surface runs perform very similarly in the round robin tournament, while the 2 isodose surface run performs poorly (Fig. 3.35c). Again, the mean modulation index, fractal dimension, and gradient entropy all show a steady increase with an increase in the number of isodose projections (Fig. 3.35d, e, f).

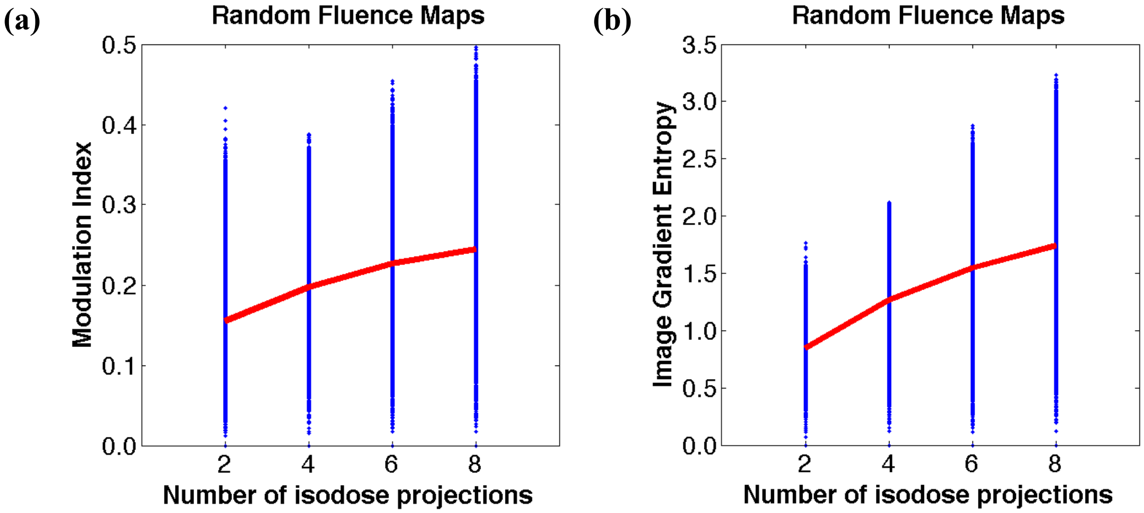

Figure 3.32: The modulation index and image gradient entropy of fluence maps of 5000 randomly generated solutions on the spine phantom with a varying number of isodose projections and no OAR reduction parameters (blue). The red curve joins the mean values.

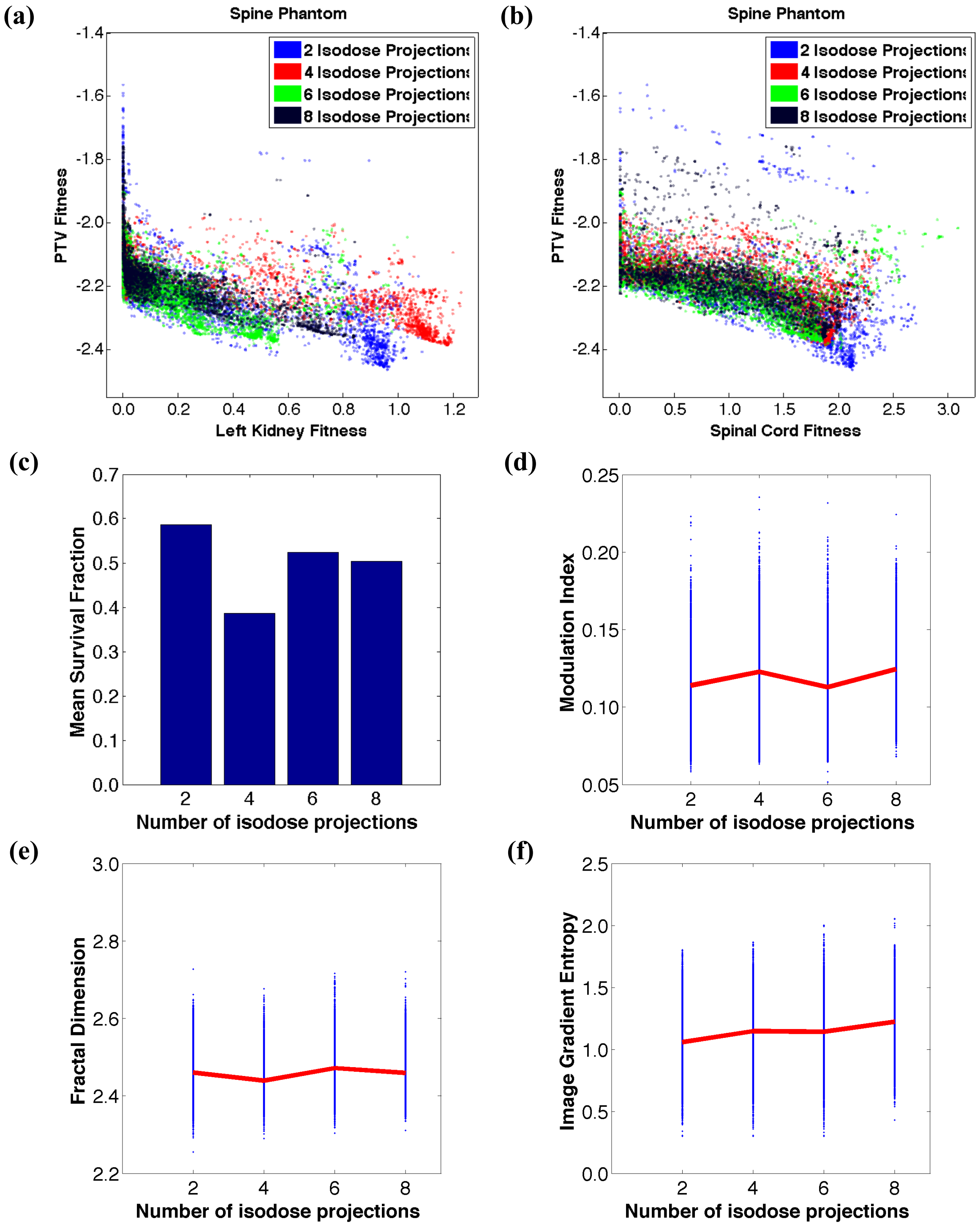

Figure 3.33: (a, b) Optimal solutions for the spine phantom from 5 beam runs done with a varying number of isodose surfaces projected onto the BEV. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

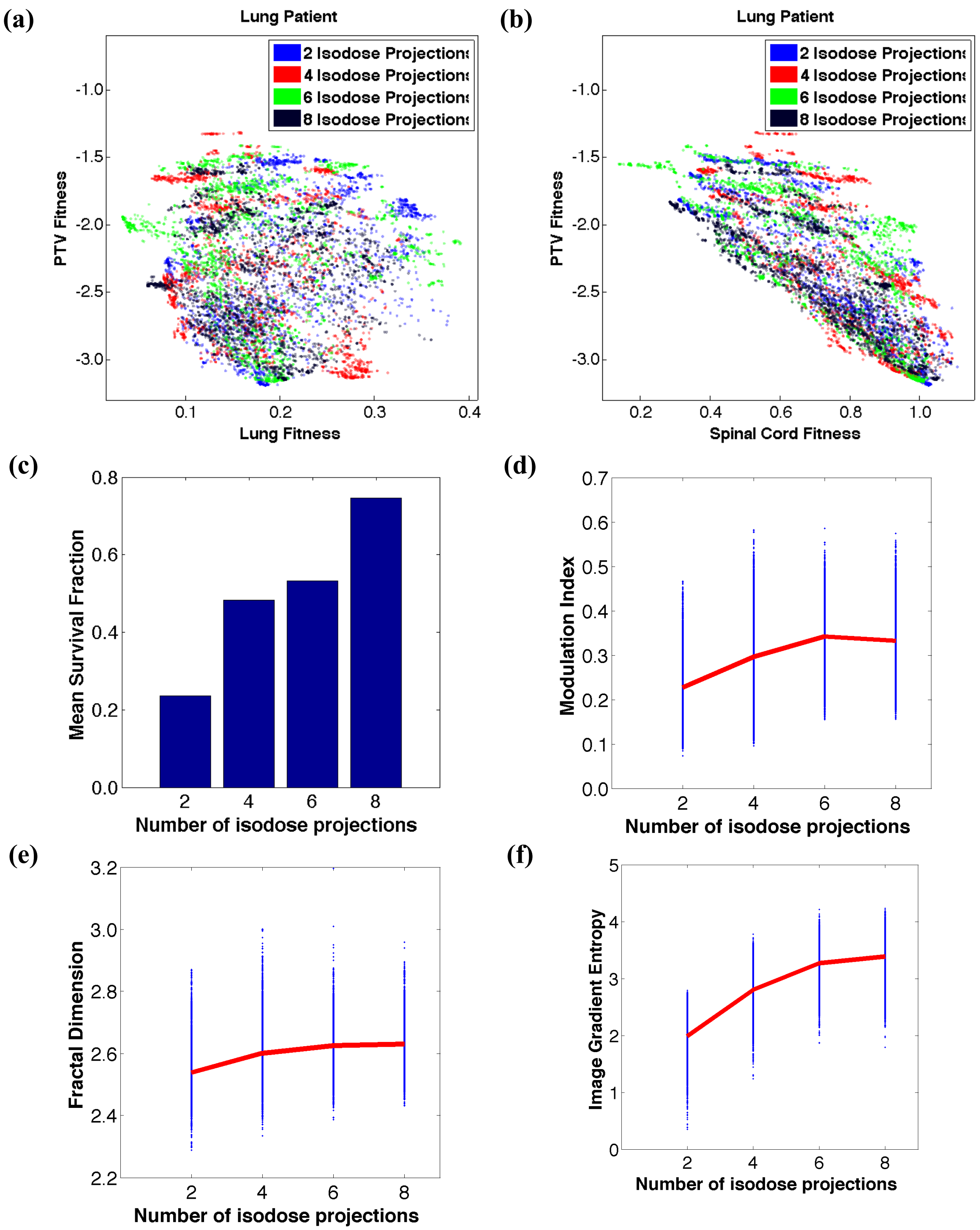

Figure 3.34: (a, b) Optimal solutions for the lung patient from 5 beam runs done with a varying number of isodose surfaces projected onto the BEV. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

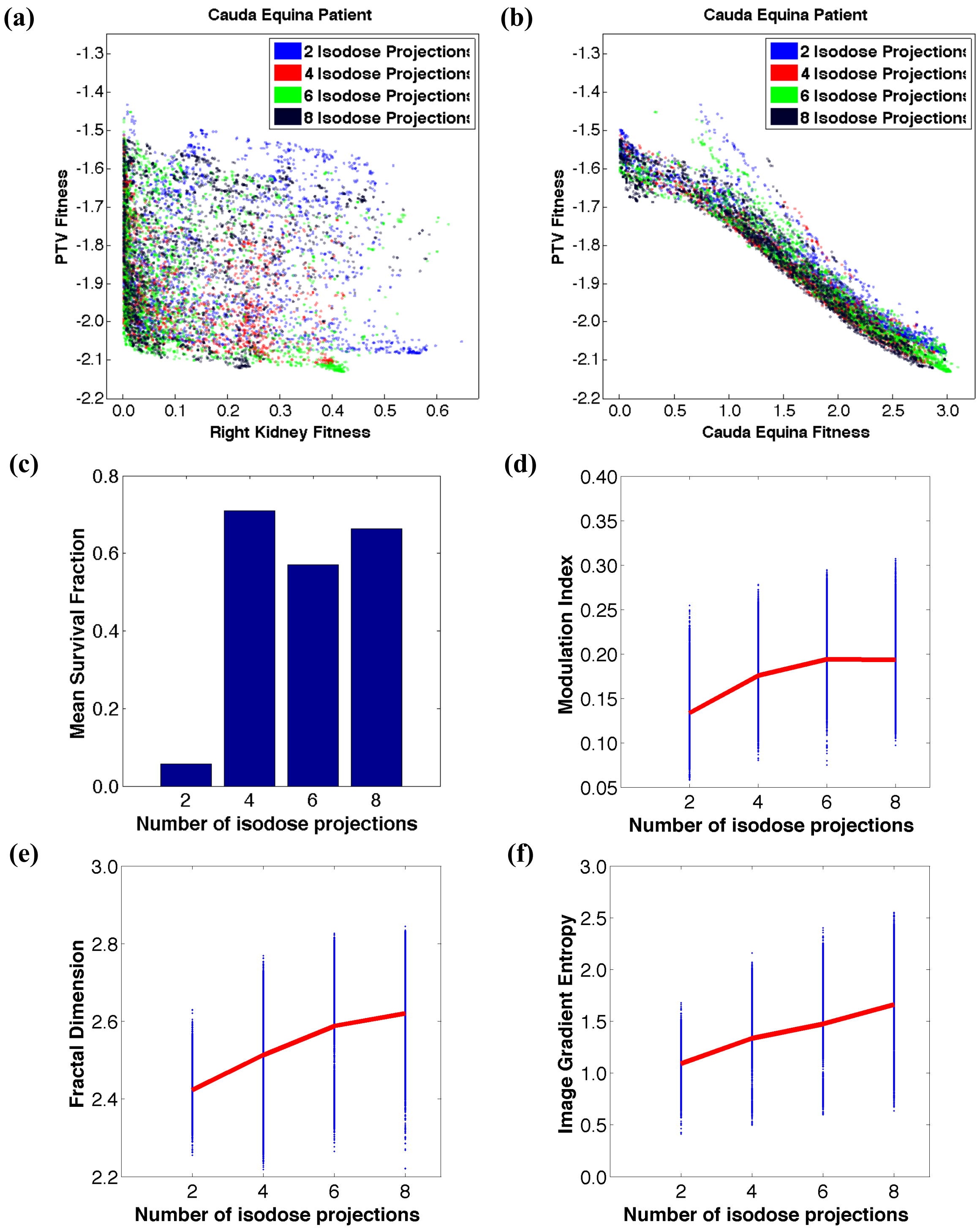

Figure 3.35: (a, b) Optimal solutions for the cauda equina patient from 5 beam runs done with a varying number of isodose surfaces projected onto the BEV. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. The winner of each game is chosen as the run with the greatest survival fraction of its solutions in a merged optimal set constructed from the pair. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red curve joins the mean value of each run.

3.3.4 Investigation into the relationship between the number of beams and fluence complexity and solution quality

The runs presented in this study have been done with five beams in order to limit the size of the search space and to place greater importance on the fluence modulation. However, in order to specifically investigate the effect of a greater number of beams on fluence complexity and solution quality, we present runs done with 11 beams and a varying number of beam groups for the lung patient (Fig. 3.27). Similar to the results of the round robin tournaments done with five beams (Fig. 3.26), the run done with 25 beam groups wins for overall solution quality, while the 81 beam group run loses (Fig. 3.27c). However, in this case the run done with 9 beam groups comes in second place. Also, although the mean modulation index still increases with the number of beam groups, the range of modulation index stays more constant than it does for the runs done with five beams (Fig. 3.27d). The behaviour of the fractal dimension and gradient entropy is very comparable to the five beam runs (Fig. 3.27e, f).

3.3.5 Comparison of PTV fluence parameterizations

An evaluation of the various parameterizations used for fluence modulation over the projection of the PTV in the BEV that are outlined in Section 3.2.2 is an important step in PARETO's evolution. In order to compare them, the best runs (as determined by round robin tournaments) from the results of Section 3.3.3 are selected for each modulation method. Thus, in this investigation, the number of parameters specified in each method is chosen so as to yield the best results for a particular patient geometry. Standard OAR reduction parameters were applied in all cases (Section 3.2.2.6). The tools used here for comparison are discussed in Section 3.2.3.

For the spine phantom, solutions from each method show much overlap in the tradeoff projections. In the tradeoff of PTV versus left kidney fitness, the linear gradient method appears to have the lowest "knee" (where solutions are simultaneously good in PTV conformity and OAR dose) and generally achieves solutions of superior PTV conformity than the other methods (Fig. 3.36a). However, in the tradeoff of PTV versus spinal cord fitness, the basic weight method achieves superior solutions in PTV conformity (Fig. 3.36b). Also, the beam group method appears to do very well at sparing the left kidney (Fig. 3.36a). Therefore, the linear gradient and basic weight methods rank closely in the round robin tournament, while the beam group method comes in third place (Fig. 3.36c). However, we note that if the margin parameter is omitted, the linear gradient method achieves solutions of superior PTV conformity compared to the basic weight method (Fig. 3.37a, b). In this case, the linear gradient method has a much higher survival fraction than the basic weight method (Fig. 3.37c). The beam group and linear gradient methods also have the largest range of modulation index and gradient entropy (Fig. 3.36d, f). The fractal dimension analysis is slightly more puzzling, showing that the mean fractal dimension of the isodose projection method is much greater than that of any other method (Fig. 3.36e).

Tradeoff projections for the lung patient again reveal much overlap between the various methods. The only significant feature is that the beam group method is able to achieve more conformal solutions in the tradeoff of PTV versus spinal cord fitness (Fig. 3.38b). However, the round robin tournament reveals that the linear gradient method has the highest mean survival fraction, while the beam group method comes in second place (Fig. 3.38c). As judged by the modulation index and gradient entropy, the beam group method is the most modulated, though the cosine transform method also has a very large range of gradient entropy values (Fig. 3.38d, f). The fractal dimension of the isodose projection method is again much higher than the other methods (Fig. 3.38e).

The results for the cauda equina patient also show that the linear gradient and beam group methods produce the highest solution quality (Fig. 3.39c). In the projection of PTV versus right kidney fitness, these methods achieve solutions with superior PTV fitness and little spread in right kidney dose (Fig. 3.39a). In the projection of PTV versus cauda equina fitness, the beam group solutions lie significantly lower in PTV fitness than most other solutions, while the linear gradient solutions are next best (Fig. 3.39b). Again, the beam group method is the most modulated (Fig. 3.39d, f) and the fractal dimension analysis is dominated by the isodose projection method (Fig. 3.39e).

Additional characteristics of the runs that are interesting to compare are the trends in mean fluence complexity visible on the tradeoff surfaces. Figures 3.40-3.45 show runs done with the linear gradient, cosine transform, and beam group methods. Each solution is colour-coded according to the mean gradient entropy (a, c, e) or modulation index (b, d, f) of its fluence maps. For the spine phantom, a general trend of increasing mean modulation index and gradient entropy with decreasing PTV fitness is noticeable, though the trend is not quite as clear for the beam group method (Fig. 3.40, Fig. 3.41). With the lung patient, the mean gradient entropy shows this trend for the linear gradient and cosine transform method (Fig. 3.42a, c Fig. 3.43a, c). The results of the beam group method show that high values of modulation index and gradient entropy occur at both low and high values of PTV fitness (Fig. 3.42e, f, Fig. 3.43e, f). However, the results for cauda equina patient show a strong trend of increasing mean modulation index and gradient entropy for decreasing PTV fitness on all tradeoff surfaces (Fig. 3.44, Fig. 3.45).

In order to directly compare the metrics employed in this work for evaluating fluence map complexity, we plot the modulation index, fractal dimension, and gradient entropy of fluence maps from solutions of varying patient geometries, fluence parameterizations, and numbers of parameters (Fig. 3.46). Figure 3.46a shows some correlation between the modulation index and gradient entropy for the basic weight, linear gradient, and isodose projection methods. Figure 3.46b shows the fractal dimension is somewhat correlated to the modulation index for the beam group and isodose projection methods. However, there is very little correlation between the fractal dimension and the gradient entropy (Fig. 3.46c).

Each PTV fluence modulation method may also be compared in terms of run time. Figure 3.47 shows the run times for each run presented in Figures 3.36-3.39. In general, the methods perform similarly, except that the isodose projection method takes much longer. Therefore, in future work we may optimize this method for speed by implementing a faster isosurface calculation algorithm.

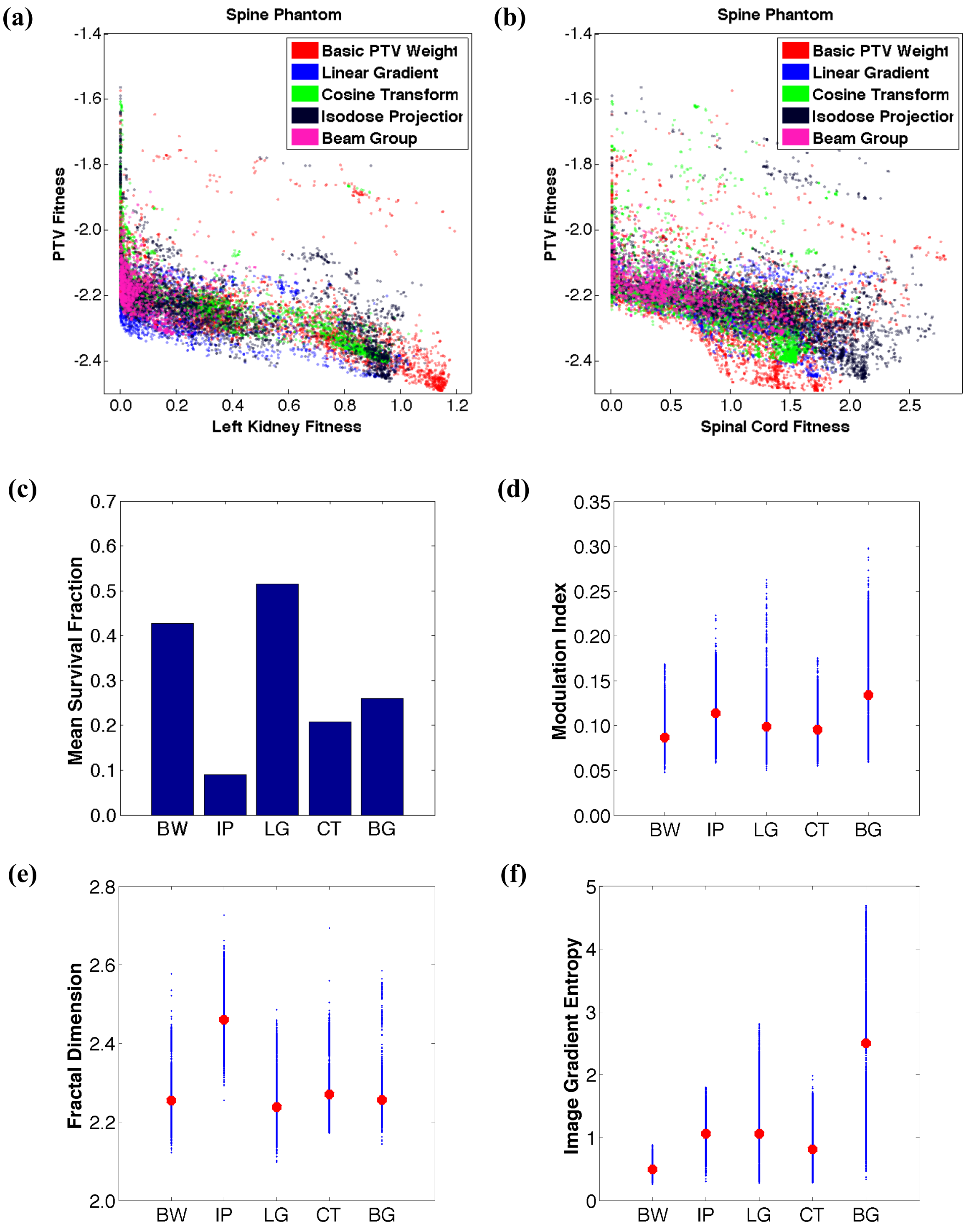

Figure 3.36: (a, b) Optimal solutions for the spine phantom from 5 beam runs done with the basic weight (BW), isodose projection (IP), linear gradient (LG), cosine transform (CT), and beam group (BG) methods with 1, 2, 3, 6, and 9 PTV fluence parameters, respectively. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red dots show the mean value of each run.

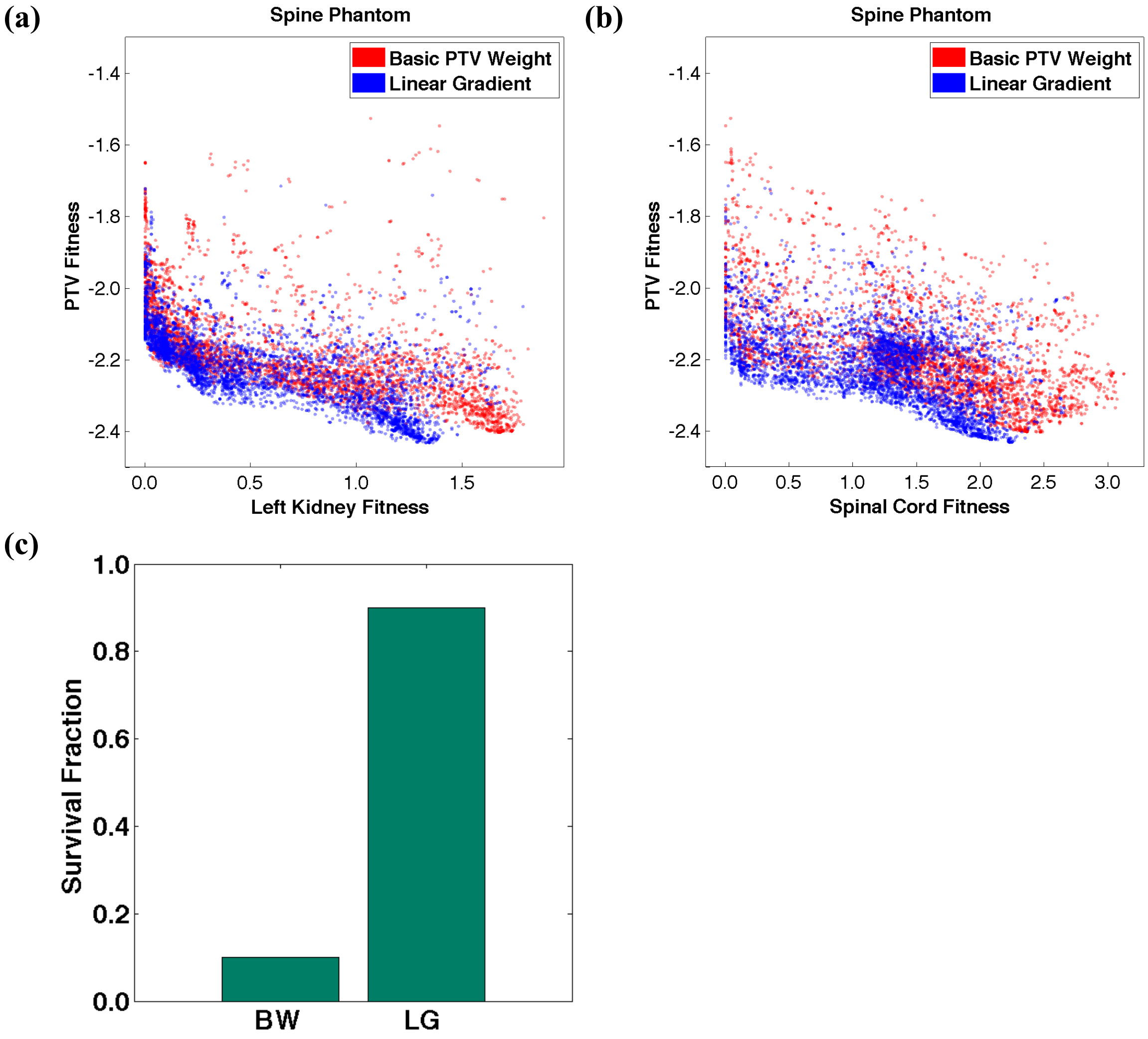

Figure 3.37: (a, b) The set of non-dominated solutions from a 5 beam run done on the spine phantom with the basic weight method (red) or the single linear gradient method (blue) with no BEV margin parameter used in either case. Tradeoff surfaces are shown for PTV conformity fitness versus OAR dose. (c) The survival fraction of each run in a set of non-dominated solutions constructed by applying Ferret's tournament selection operator to the combination of all solutions from both runs.

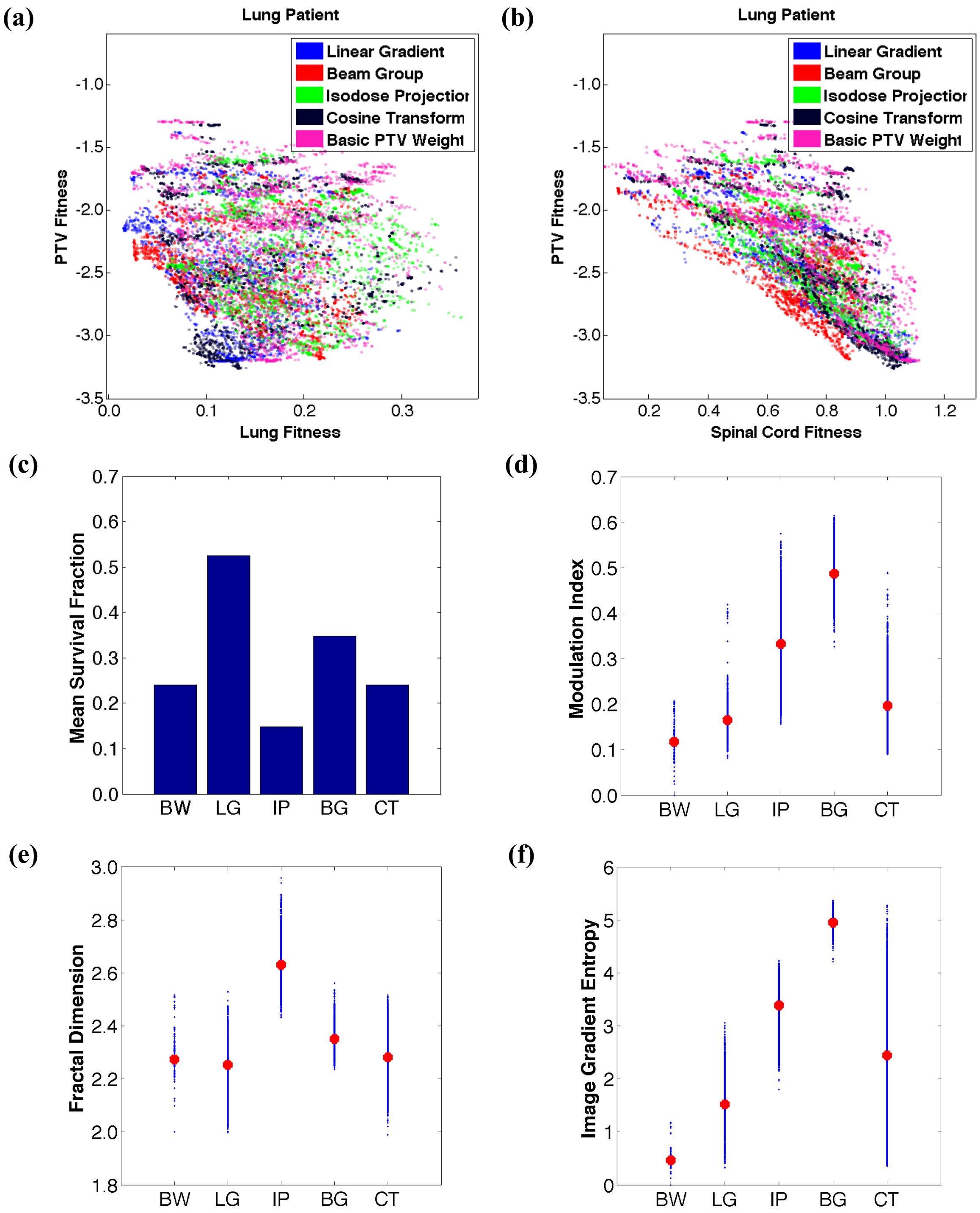

Figure 3.38: (a, b) Optimal solutions for the lung patient from 5 beam runs done with the basic weight (BW), linear gradient (LG), isodose projection (IP), beam group (BG), and cosine transform (CT) methods with 1, 6, 8, 49, and 58 PTV fluence parameters, respectively. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red dots show the mean value of each run.

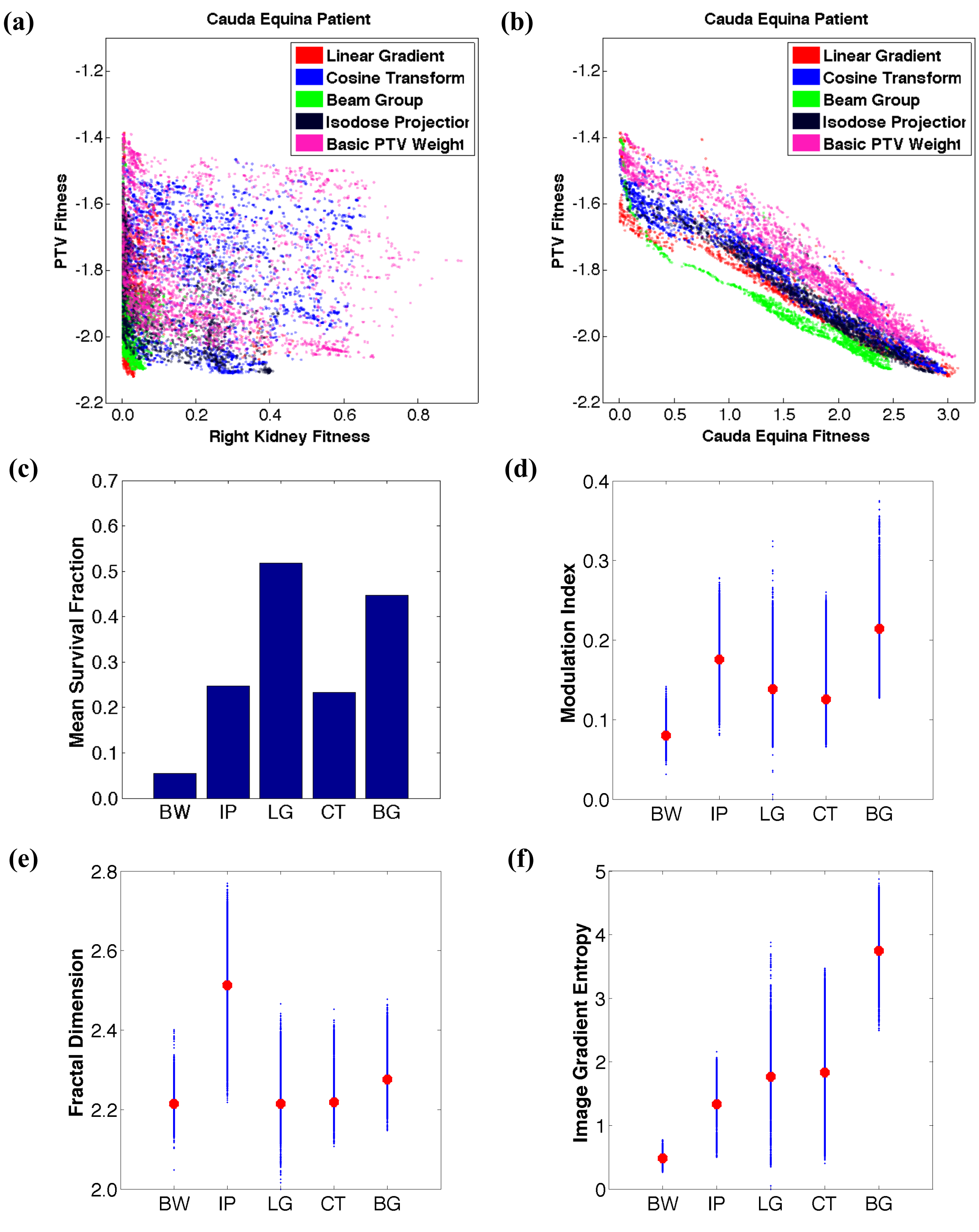

Figure 3.39: (a, b) Optimal solutions for the cauda equina patient from 5 beam runs done with the basic weight (BW), isodose projection (IP), linear gradient (LG), cosine transform (CT), and beam group (BG) methods with 1, 4, 6, 17, and 25 PTV fluence parameters, respectively. Various tradeoff surfaces of PTV fitness versus OAR dose are shown. (c) The mean survival fraction of each run from a round robin tournament in which each run plays every other run. (d, e, f) The modulation index, fractal dimension, and image gradient entropy of all fluence maps of all solutions of each run are shown (blue). The red dots show the mean value of each run.

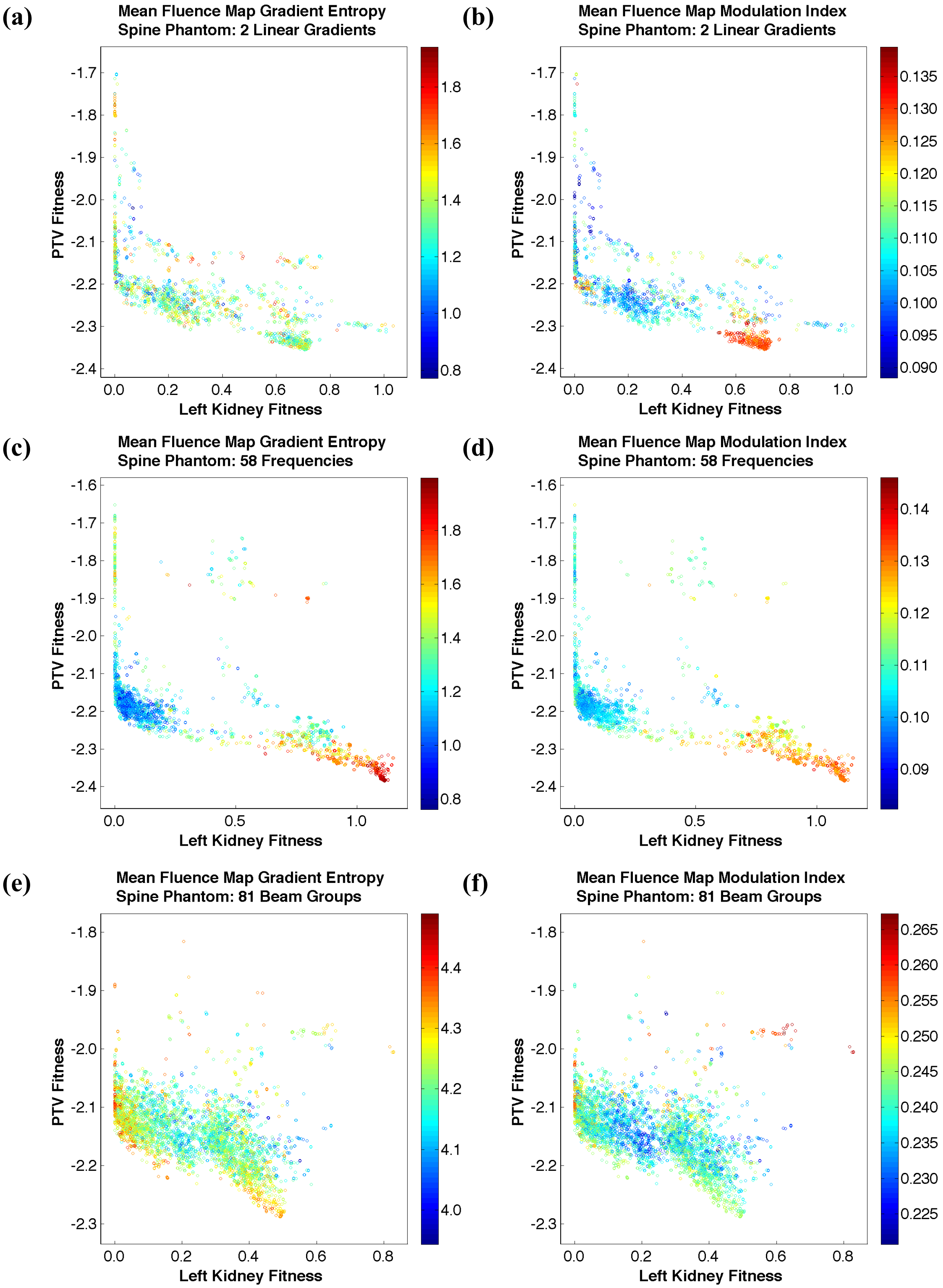

Figure 3.40: The mean image gradient entropy (left) and modulation index (right) of each optimal plan for the spine phantom projected into the plan of PTV fitness versus left kidney fitness for a 5 beam run done with 2 linear gradients (a,b), 58 frequency parameters (c,d), and 81 beam groups (e,f).

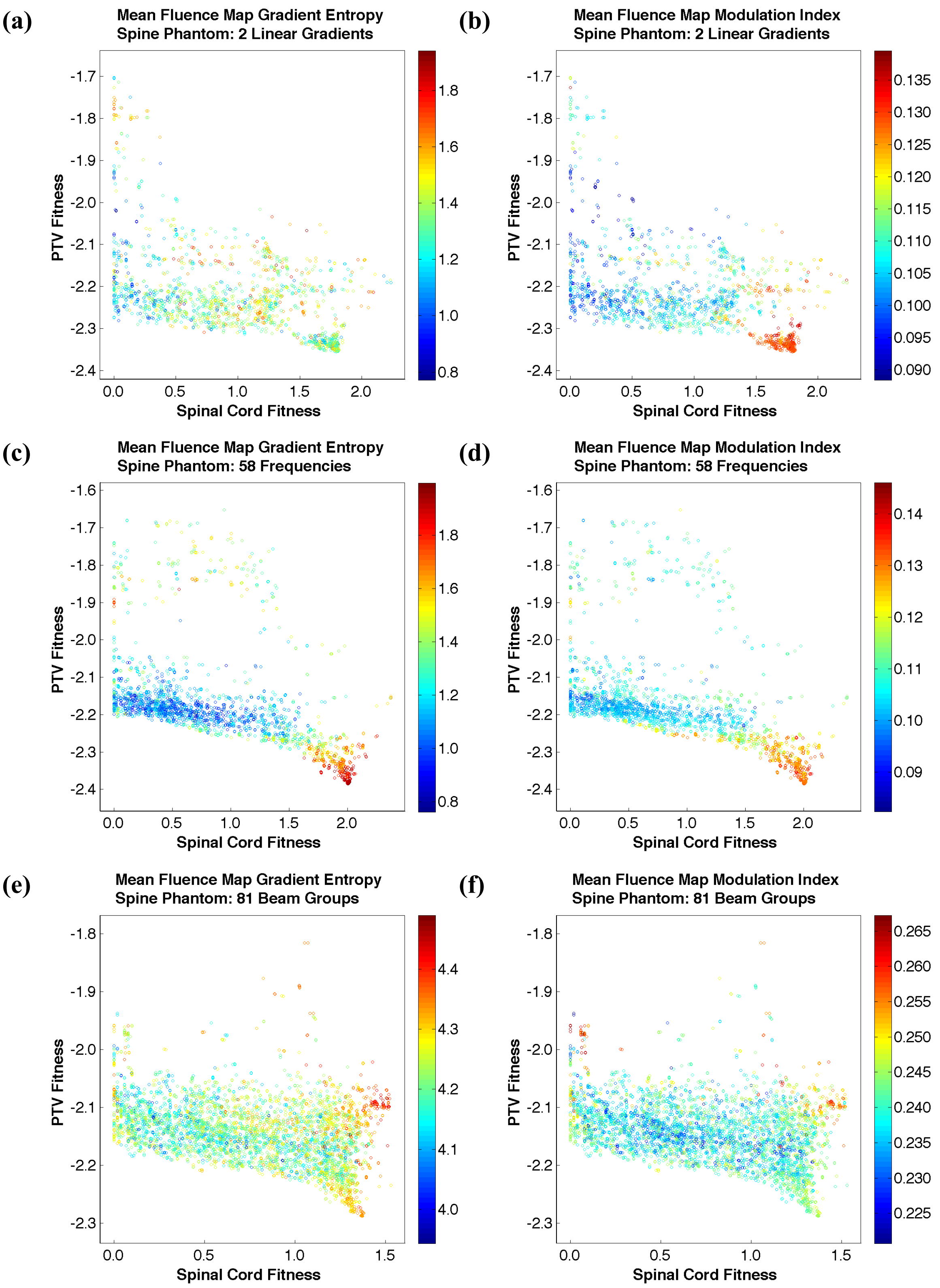

Figure 3.41: The mean image gradient entropy (left) and modulation index (right) of each optimal plan for the spine phantom projected into the plan of PTV fitness versus spinal cord fitness for a 5 beam run done with 2 linear gradients (a,b), 58 frequency parameters (c,d), and 81 beam groups (e,f).

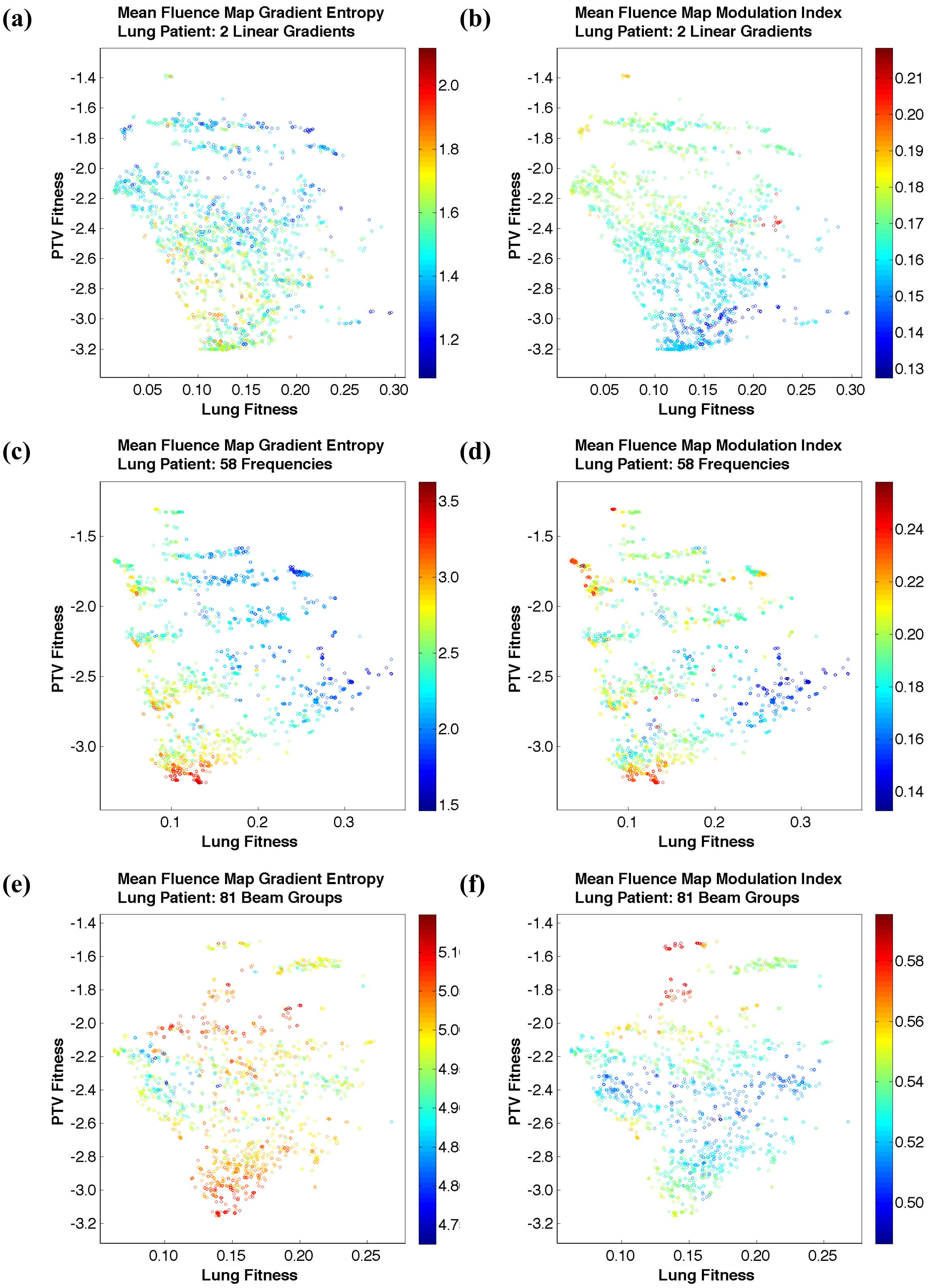

Figure 3.42: The mean image gradient entropy (left) and modulation index (right) of each optimal plan for the lung patient projected into the plan of PTV fitness versus lung fitness for a 5 beam run done with 2 linear gradients (a,b), 58 frequency parameters (c,d), and 81 beam groups (e,f).

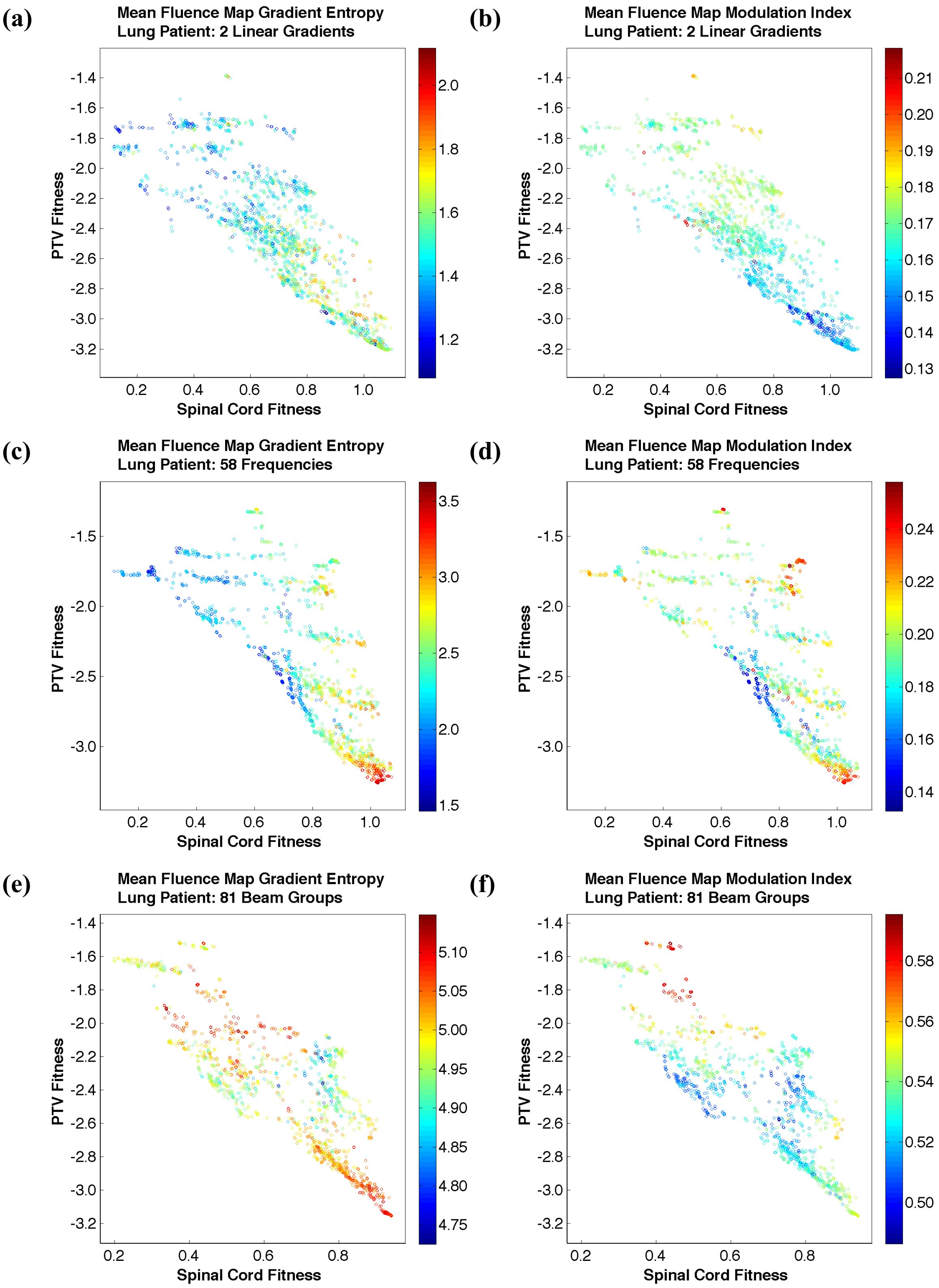

Figure 3.43: The mean image gradient entropy (left) and modulation index (right) of each optimal plan for the lung patient projected into the plan of PTV fitness versus spinal cord fitness for a 5 beam run done with 2 linear gradients (a,b), 58 frequency parameters (c,d), and 81 beam groups (e,f).

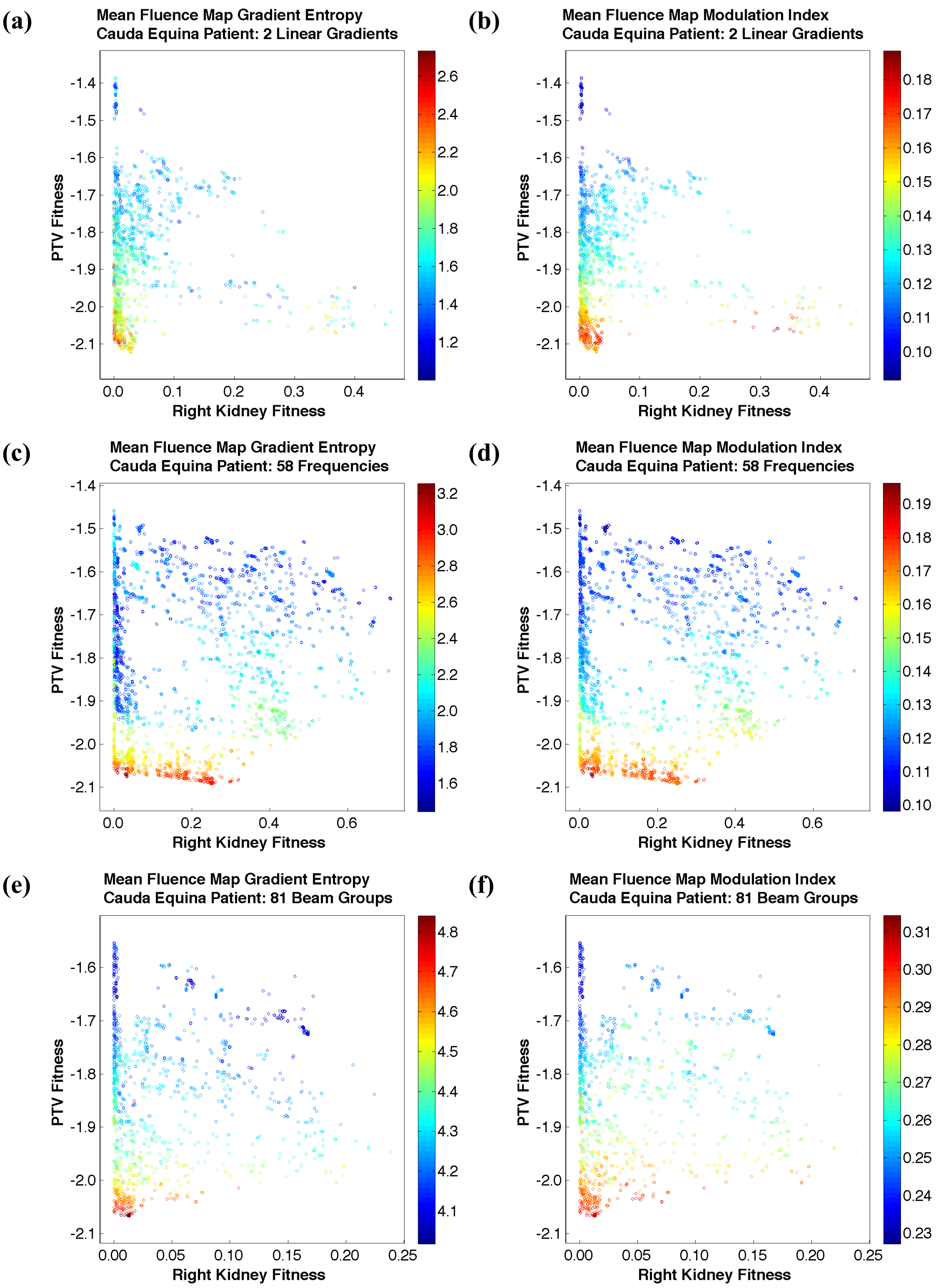

Figure 3.44: The mean image gradient entropy (left) and modulation index (right) of each optimal plan for the cauda equina patient projected into the plan of PTV fitness versus right kidney fitness for a 5 beam run done with 2 linear gradients (a,b), 58 frequency parameters (c,d), and 81 beam groups (e,f).

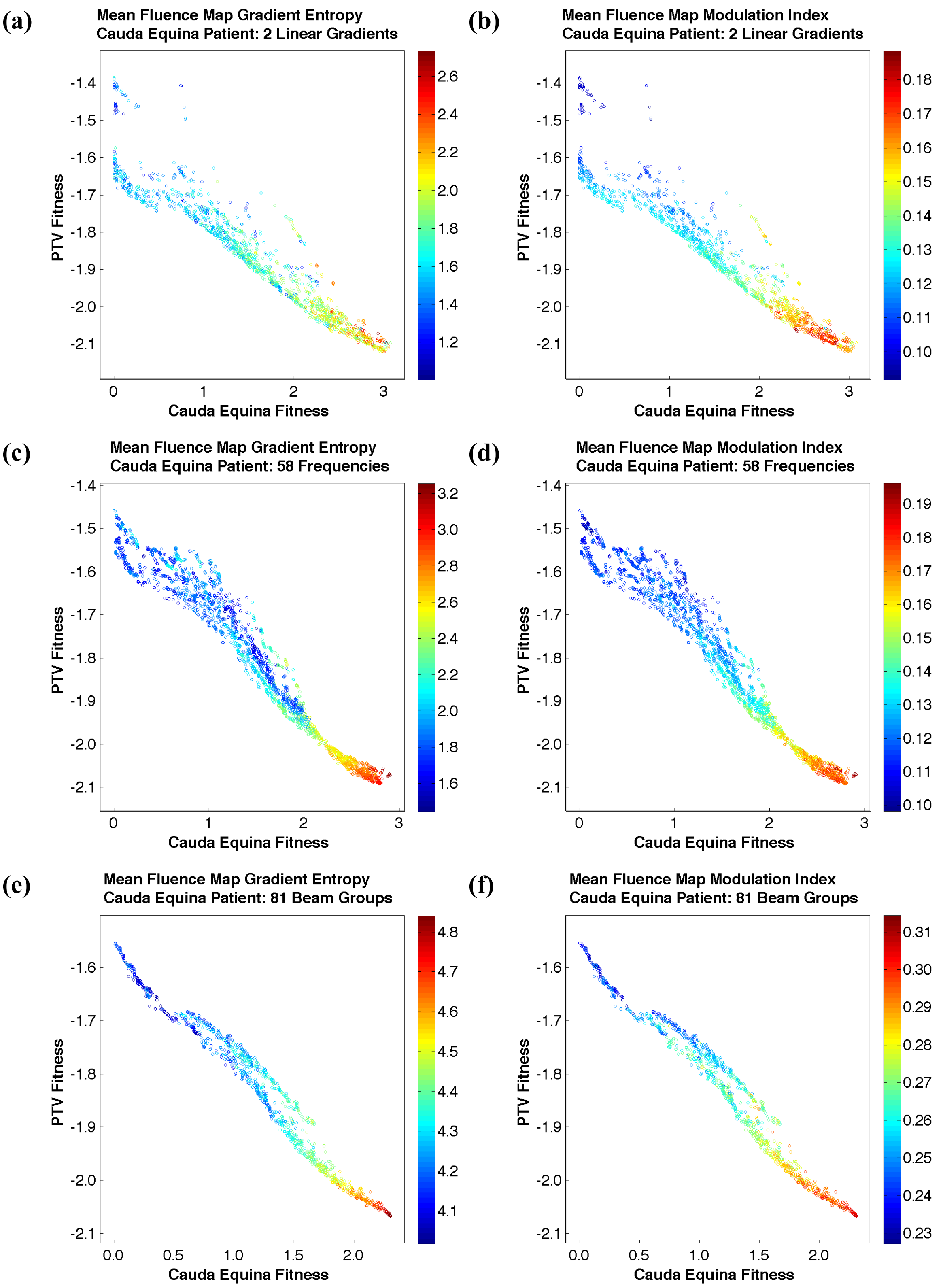

Figure 3.45: The mean image gradient entropy (left) and modulation index (right) of each optimal plan for the cauda equina patient projected into the plan of PTV fitness versus cauda equina fitness for a 5 beam run done with 2 linear gradients (a,b), 58 frequency parameters (c,d), and 81 beam groups (e,f).

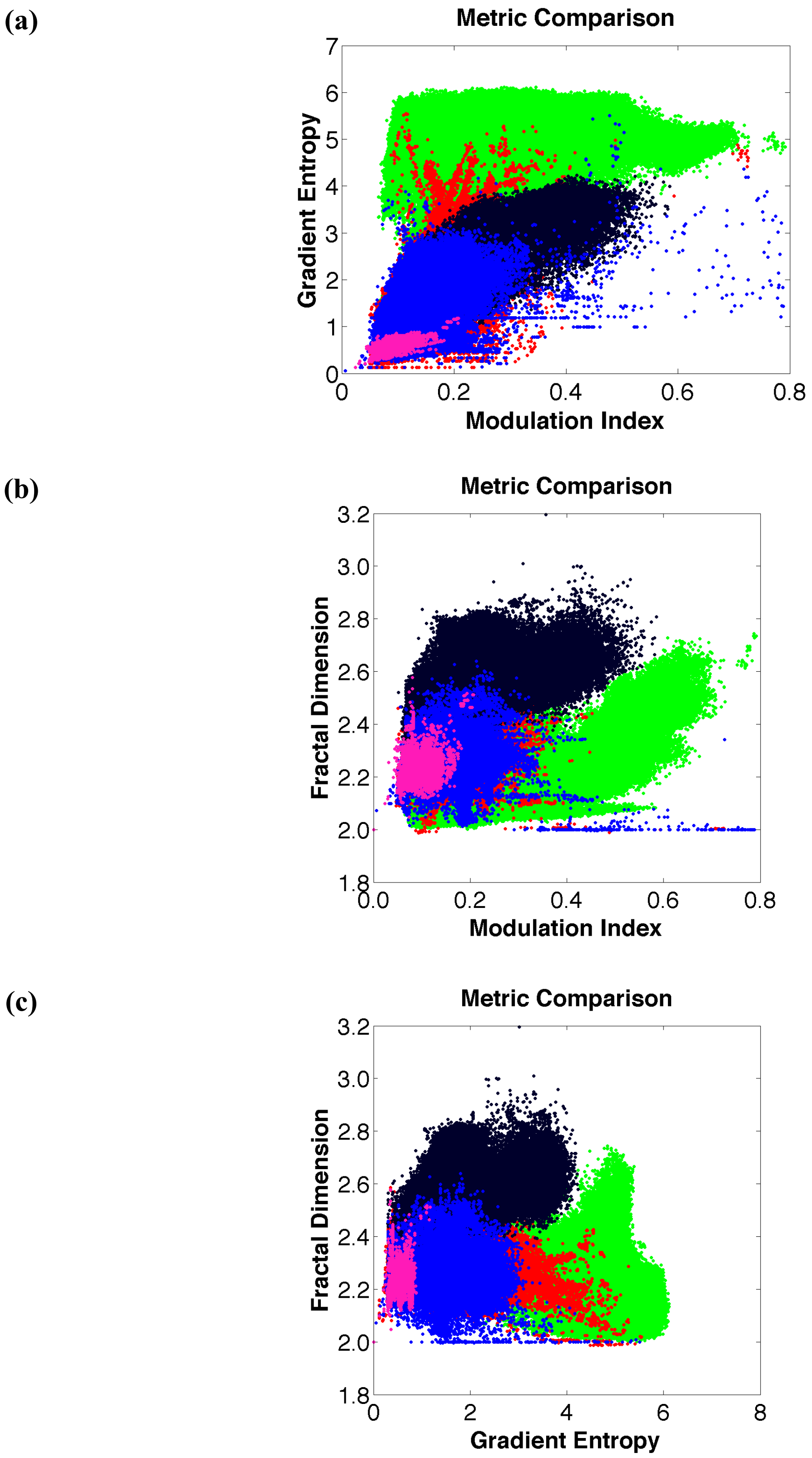

Figure 3.46: Fluence map metrics from all solutions of runs done on all geometries with the basic weight (pink), linear gradient (blue), cosine transform (red), beam group (green), and isodose projection (black) methods.

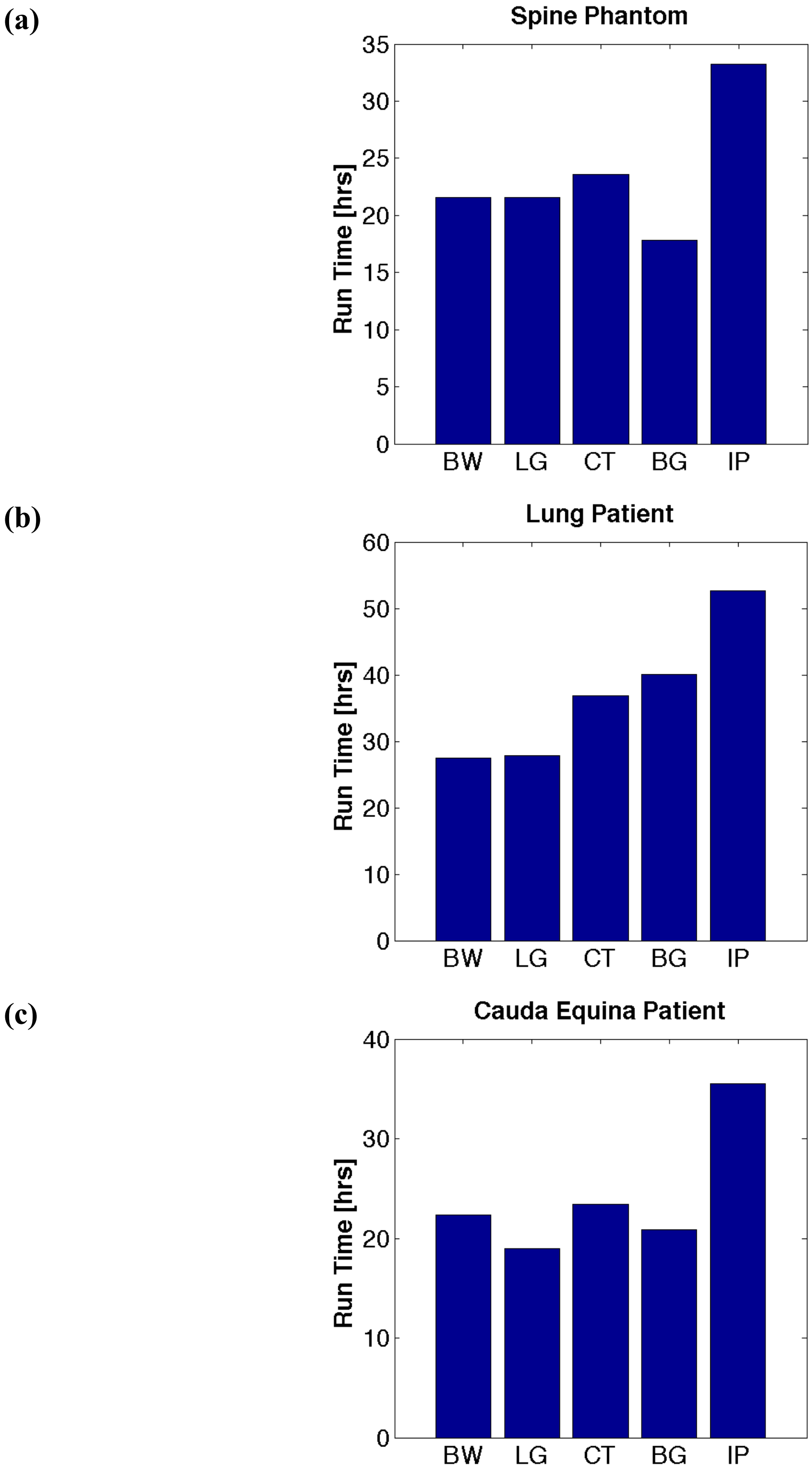

Figure 3.47: Run times for the basic weight (BW), linear gradient (LG), cosine transform (CT)*, beam group (BG), and isodose projection (IP) methods compared in Figures 3.36-3.39.

***Note: the CT run for the spine phantom was performed using 8 CPUs, whereas all other runs were performed using 12 CPUs.**

3.3.6 Investigation into the effect of applying fluence gradients onto the projections of the OARs that overlap the PTV in the BEV

The major focus of this chapter is to evaluate the merits of various parameterizations of fluence modulation for the projection of the PTV in the BEV. OARs that overlay the PTV projection are typically handled by simply multiplying the PTV fluence by a reduction parameter, which preserves the character of the underlying gradient but effectively spares the OARs (see Section 3.2.2.6). However, in this section, we investigate the effect of multiplying a single linear gradient function (specified by Equation 3.4 and 3.5) by the PTV fluence over each OAR projection. For each patient geometry, a single linear gradient was used for PTV modulation on a run done with OAR reduction parameters and a run done with linear gradients multiplied over the OAR projections.

For the spine phantom, OAR gradients do not improve solution quality. In the tradeoff of PTV versus left kidney fitness, solutions from the run done with OAR reduction parameters are much lower in PTV fitness (Fig. 3.48b). Therefore, in the merged optimal set constructed from the solutions of both runs, the run done with OAR reduction parameters has a much higher survival fraction (Fig. 3.48c).

For both the lung and cauda equina patients, however, OAR gradients are very useful. In the tradeoff of PTV versus lung fitness, the run done with a linear gradient on the OARs produces solutions of much greater PTV conformity (Fig. 3.49b). Thus, the fraction of solutions that survive in the merged set is much greater than for the run done with OAR reduction parameters (Fig. 3.49c). Similarly, the solutions of the run done with OAR gradients are generally much lower in PTV fitness for the cauda equina patient (Fig. 3.50a, b) and have a higher survival fraction in the merged set (Fig. 3.50c).

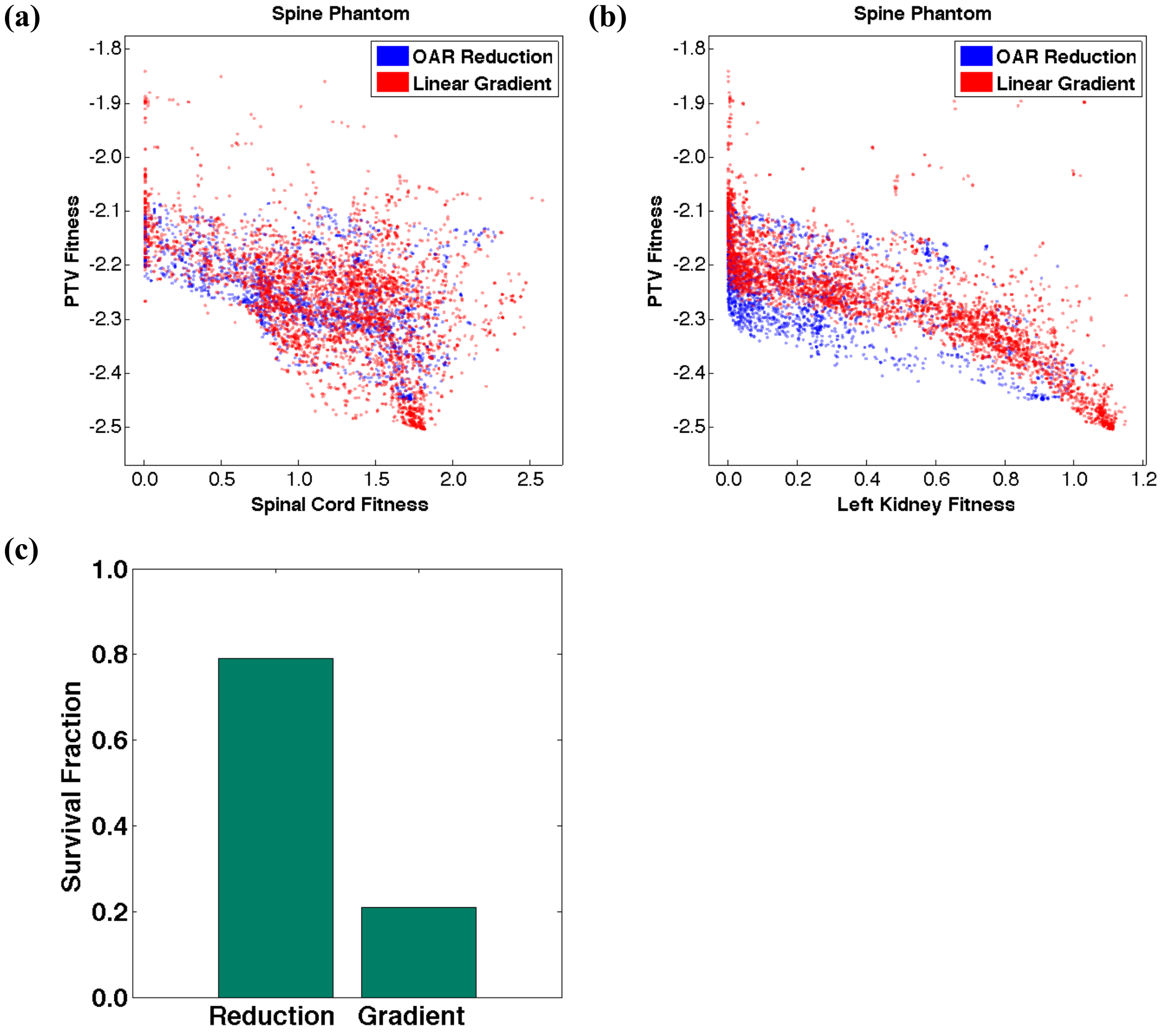

Figure 3.48: (a, b) The set of non-dominated solutions from a 5 beam run done on the spine phantom with a single linear gradient for PTV modulation and OAR reduction parameters (blue) or a single linear gradient on each OAR (red). Tradeoff surfaces are shown for PTV conformity fitness versus OAR dose. (c) The survival fraction of each run in a set of non-dominated solutions constructed by applying Ferret's tournament selection operator to the combination of all solutions from both runs.

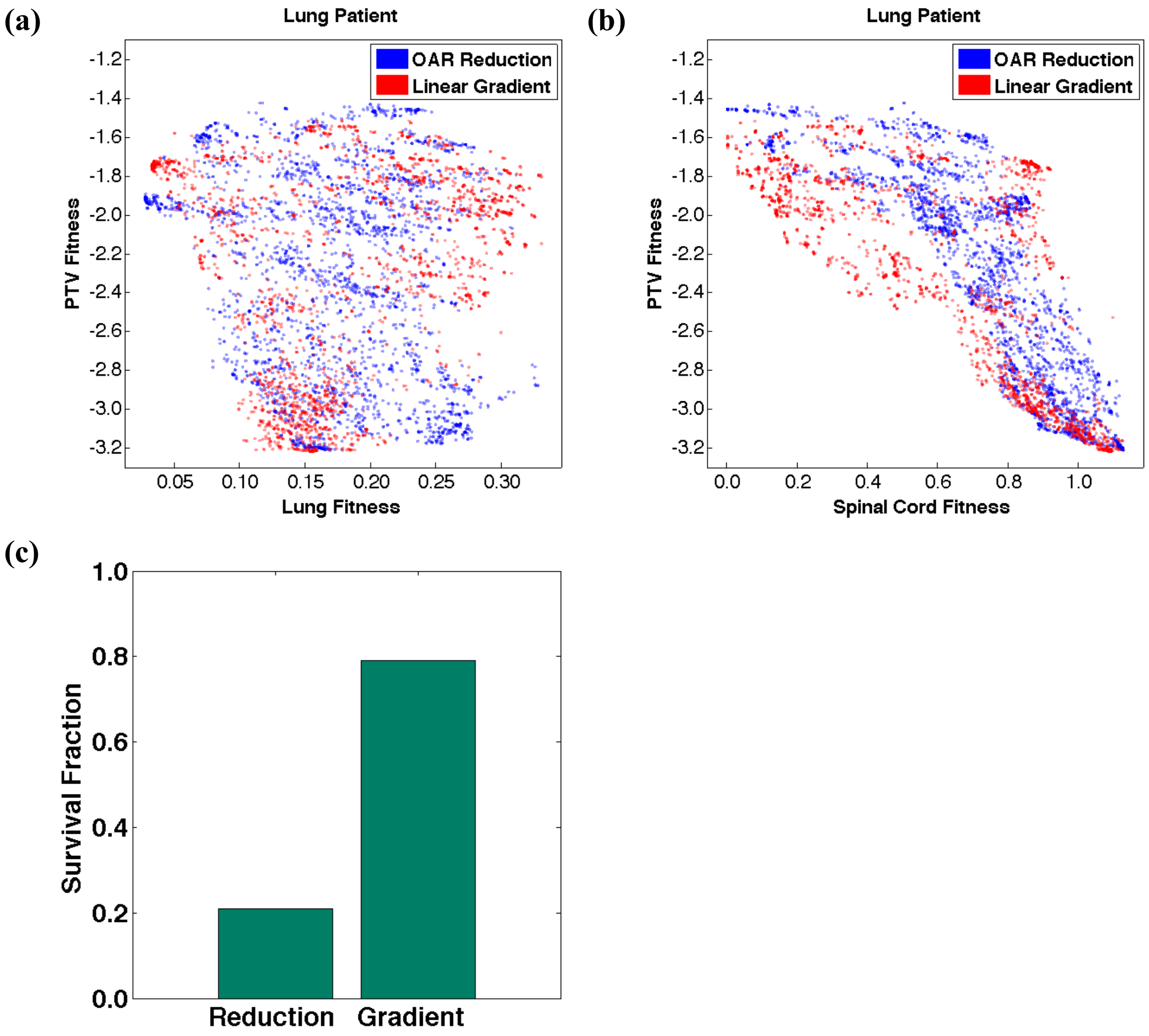

Figure 3.49: (a, b) The set of non-dominated solutions from a 5 beam run done on the lung patient with a single linear gradient for PTV modulation and OAR reduction parameters (blue) or a single linear gradient on each OAR (red). Tradeoff surfaces are shown for PTV conformity fitness versus OAR dose. (c) The survival fraction of each run in a set of non-dominated solutions constructed by applying Ferret's tournament selection operator to the combination of all solutions from both runs..

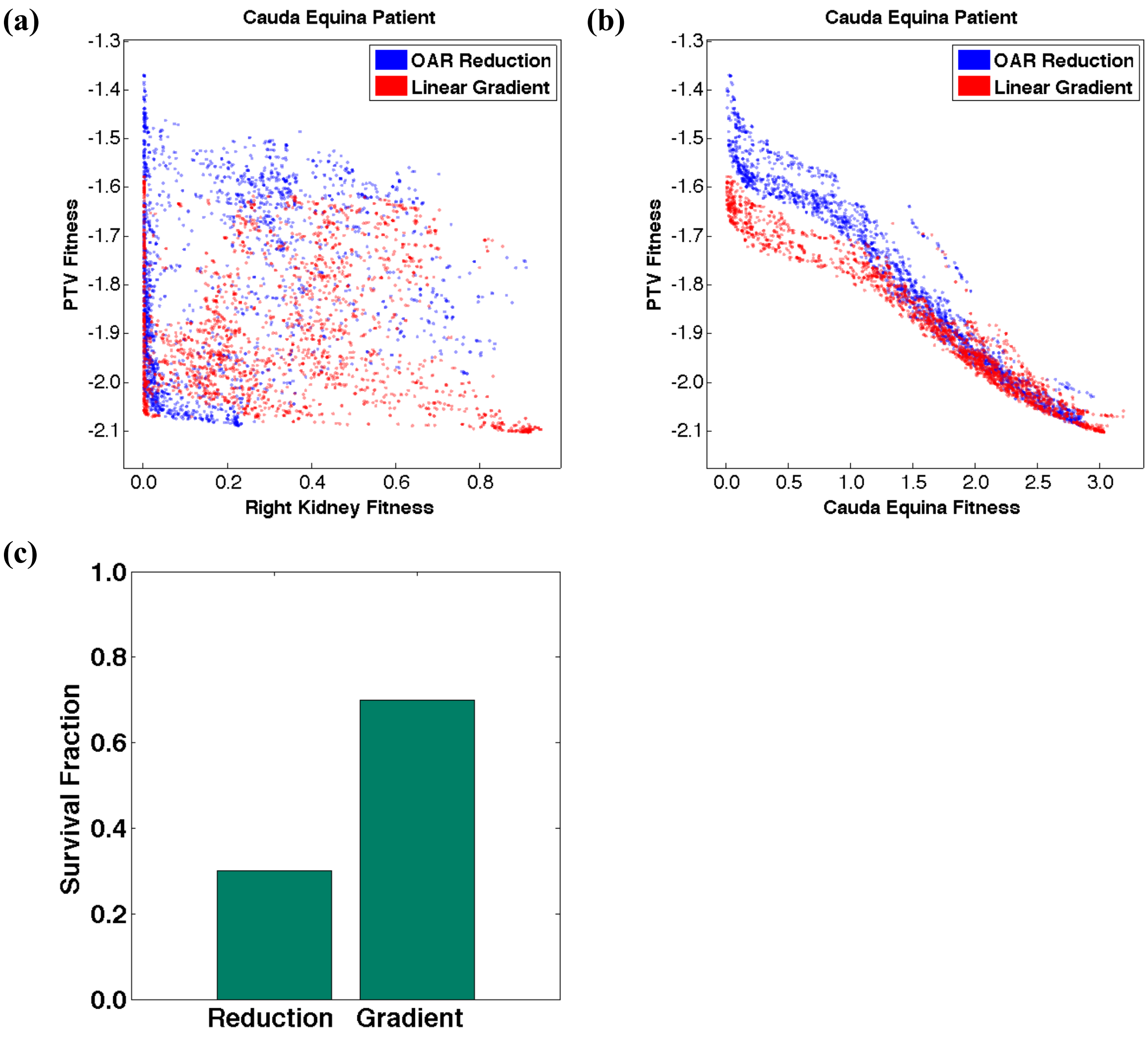

Figure 3.50: (a, b) The set of non-dominated solutions from a 5 beam run done on the cauda equina patient with a single linear gradient for PTV modulation and OAR reduction parameters (blue) or a single linear gradient on each OAR (red). Tradeoff surfaces are shown for PTV conformity fitness versus OAR dose. (c) The survival fraction of each run in a set of non-dominated solutions constructed by applying Ferret's tournament selection operator to the combination of all solutions from both runs.

3.4 Discussion

3.4.1 Demonstration of the effect of an optimized BEV margin parameter on the spine phantom

The results of Section 3.3.1 show that an optimized BEV margin parameter is extremely useful in improving solution quality. The survival fraction of spine phantom solutions for a run done with the margin parameter (SF=96%) is much higher than the run done without (SF=4%). Therefore, all of the runs presented in this study have been done with an optimized margin parameter for each beam.

The spine phantom represents a common patient geometry encountered in radiotherapy. When radiotherapy is delivered by static beams or arc therapy, a homogeneous dose distribution in a concave PTV requires a steep fluence gradient towards the concavity [30]. This can be explained by visualizing many beams that are oriented tangential to the OAR with a fixed beam width (Fig. 1.2). In this case, fewer beams pass through a point closer to the concavity than one farther away. However, modifying beam width is an effective means to create a more uniform BIV (Fig. 3.9), which leads to a more homogeneous dose distribution. In our implementation, dilating or eroding the projection of the PTV in the BEV modifies the effective beam width. Therefore, it is possible that little additional PTV modulation is needed if the dose inhomogeneity usually caused by the concave geometry is greatly reduced by the margin parameter and OARs are spared with OAR reduction parameters. While OAR sparing causes greater PTV inhomogeneity, using a modest number of beams (approximately five) may be enough to accommodate for this. Our results show that using only a beam weight, a margin parameter, and OAR reduction parameters is a very effective method for the spine phantom (see Section 3.3.1). Also, for all of the parameterizations investigated on the spine phantom, we have seen that increasing the number of parameters in order to increase the fluence complexity degrades solution quality when OAR reduction parameters are used (see Section 3.3.3). However, we note that increasing the number of parameters does not always lead to a clear increase in fluence complexity.

3.4.2 Comparison of the coupled and decoupled parameterizations of the linear gradient method

The results of Section 3.3.2 show that both the coupled and decoupled parameterizations of the linear gradient method are capable of producing high quality solutions that lie in a similar region of objective function space. The merged optimal sets from a run performed with each method show that each parameterization obtains the greatest survival fraction at least once on both patient geometries (3.11e, 3.12e). Thus, we see that there is no real advantage of one method over the other. In some cases the solutions generated by each method occupy different regions of objective function space in the merged optimal sets (Fig. 3.11c, d, Fig. 3.12c, d), but the trends do not hold for all patient geometries. Therefore, in the other investigations presented in this work, we have chosen to use the decoupled parameterization simply because it is more intuitive.

3.4.3 Investigation into the effect of varying the number of PTV fluence parameters on fluence complexity and solution quality

3.4.3.1 Linear Gradient Method

As discussed in Section 3.2.2.2, with our implementation of the linear gradient method, we expect a decrease in fluence complexity for a large number of multiplied gradient maps. Indeed, randomly generated fluence maps show that the mean modulation index and gradient entropy decrease past two multiplied gradients (see Section 3.3.3.1). For the spine phantom and lung patient, increasing the number of linear gradients that are multiplied together for PTV modulation causes a modest increase in the mean modulation index and gradient entropy of the fluence maps. However, there is no clear trend for the cauda equina patient unless the modulation index and gradient entropy are calculated prior to the application of OAR reduction parameters. Therefore, the increase in modulation gained by multiplying a number of gradients together is fairly small considering that multiplication by OAR reduction parameters (causing more sharp edges on the maps) can destroy the trend.

For the spine phantom, using higher numbers of gradients results in poorer solution quality. This supports the notion that a beam weight, an optimized BEV margin parameter, and OAR reduction parameters may be sufficient for fluence modulation with the spine phantom (Section 3.4.1). Some results for the lung and cauda equina patients show that an increase in the number of gradients is beneficial (Fig. 3.15, 3.17).

Nonetheless, when we repeat the runs done on the lung patient, we see no reliable improvement in solution quality for larger numbers of gradients (Fig. 3.16). This may be because the range of the modulation index does not always increase with the number of gradients as much as it does for the run shown in Figure 3.15d (although the mean value does increase). Also, the relative increase in fluence complexity gained by adding more parameters is not nearly as great for the linear gradient method as it is for other methods such as the beam group method (see Fig. 3.24d, 3.25d, 3.28d).

3.4.3.2 Cosine Transform Method

Adding higher frequency components is intended to increase the complexity of the fluence maps produced by the cosine transform method. For randomly generated fluence maps, the mean modulation index jumps significantly at a large number of frequencies, while the mean gradient entropy jumps at a smaller number (see Section 3.3.3.2). For PARETO solutions, the gradient entropy is generally the most sensitive to a change in the number of frequency parameters. The modulation index, on the other hand, saturates quickly with the number of parameters. Therefore, there is no stable trend of an increase in the amount of steep gradients on the maps (as judged by the modulation index) as higher frequency components are added. However, there is a greater spread in the strength of the gradients on the map (as judged by the gradient entropy). This behaviour is not apparent with other fluence parameterizations. For an evaluation of the fluence metrics used, see Section 3.4.7.

For the cosine transform method, trends in solution quality are approximately consistent with other methods. For the spine phantom, there is generally a decrease in solution quality with an increase in the number of frequency parameters. With the lung patient, using a very large number of frequency parameters improves solution quality considerably. Similarly, a moderate to high number of frequency parameters is appropriate for the cauda equina patient.

3.4.3.3 Beam Group Method

When the resolution of the beam group grid is increased, the amount of fluence modulation between adjacent pixels is directly increased. The modulation index and gradient entropy show this trend on randomly generated maps for all patient geometries, while the fractal dimension concurs in most cases (see Section 3.3.3.3). Still, the gradient entropy increases less steeply at higher numbers of beam groups, showing that very high grid resolutions have a larger impact on the typical magnitude of variation between adjacent pixels (characterized by the modulation index) than on the spread of pixel variations (characterized by the gradient entropy).

Increasing the number of beam groups on the spine phantom clearly degrades solution quality (Fig. 3.24c). This finding is consistent with results from other methods and is also strongly supported by reproducibility tests on beam group runs (Fig. 3.26). With the lung patient, however, solution quality gradually improves up to a moderate number of beam groups and then begins to degrade. Results for the cauda equina patient show a strong improvement in solution quality for 25 beam groups, but little difference between the other runs. Therefore, in an effort to see a more gradual improvement in solution quality followed by a saturation or degradation as the number of beam groups is increased for both the spine phantom and cauda equina patient, we performed runs that omitted OAR reduction parameters. However, the main effect was to simply degrade the quality of the run done with a very small number of beam groups. Nonetheless, this is one of the only situations that we have found where increasing the amount of fluence modulation is helpful on the spine phantom, demonstrating that OAR reduction parameters and an optimized BEV margin parameter are the main factors responsible for determining its solution quality.

3.4.3.4 Isodose-based Projection Method

Finally, for the isodose projection method, the number of isodose surfaces that are calculated and projected onto each BEV determines the number of distinct regions that are assigned various weights on a fluence map. For the lung and cauda equina patients, the modulation index, gradient entropy, and fractal dimension all increase with the number of isodose projections (see Section 3.3.3.4). For the spine phantom PARETO solutions, there is little variation in the amount of modulation with the number of isodose projections. However, the random spine phantom solutions (with no OAR reduction parameters applied) do show an increase in the mean modulation index and gradient entropy for a greater number of isodose projections. Therefore, the GA appears to have chosen to suppress the greater fluence complexity possible with a larger number of isodose projections by choosing fairly similar weight values to apply to the projections, leading to flatter fields.

For the spine phantom, there is no obvious trend in solution quality with the number of parameters due to the small variation in the amount fluence modulation between the runs. However, the least modulated run (done with two isodose projections) still turns out to be the best, which is consistent with other methods. On the other hand, solution quality does steadily increase with the number of isodose projections for the lung patient. For the cauda equina patient, a moderate number of isodose projections results in high quality solutions, but a small number performs poorly.

Thus, we see that for most PTV modulation methods, there is no gradual increase in solution quality for an increase in fluence complexity. There are only two significant cases where this occurs: for a varying number of beam groups and a varying number of isodose projections on the lung patient (Fig. 3.25, 3.34). For the beam group method, this result is reproducible (see Fig. 3.26). However, due to long run times, we have not reproduced the results of most other methods, and thus the tradeoff surfaces may be subject to stochastic variation between runs. Still, we have seen that for the lung and cauda equina patients, runs done with moderate or highly modulated fields always outperform those done with very little modulation, while the reverse is generally true for the spine phantom.

3.4.4 Investigation into the relationship between the number of beams and fluence complexity and solution quality

Using a smaller number of beams causes a loss of target dose homogeneity [30]. Therefore, greater fluence complexity is most useful in this regime. This is supported by the fact that a run done with a small amount of fluence modulation (only nine beam groups) comes in last or second last place when five beams are used (Fig. 3.26), but does much better (second place) when eleven beams are used (Fig. 3.27c). Also, the trend of an increase in modulation index for an increase in the number of beam groups is strongest for the five beam runs (Fig. 3.25d, 3.27d). Therefore, it is preferable to work with a smaller number of beams when investigating fluence parameterizations so that the GA has more incentive to produce solutions of greater fluence complexity, pushing each parameterization to its fullest potential and thus maximizing the trends detectable.

3.4.5 Comparison of PTV Fluence Parameterizations

Each method presented in Section 3.2.2 has been compared in terms of fluence complexity, solution quality, and run time (see Section 3.3.5). For the spine phantom, the basic weight and linear gradient methods result in the highest solution quality and produce some of the least modulated fluence maps. When the BEV margin parameter is omitted, the linear gradient method (which produces slightly more modulated fluence maps overall than the basic weight method) far surpasses the basic weight method. Therefore, as discussed in Section 3.4.1, the optimized BEV margin parameter is largely responsible for achieving high solution quality on the spine phantom. The best PTV fluence parameterization for the lung patient is the linear gradient method using two multiplied gradients. Although the beam group solutions are more modulated overall and appear to have some advantages in PTV conformity, solutions from the linear gradient method have the highest survival fraction in the round robin tournament. Similarly, for the cauda equina patient the round robin tournament shows that the linear gradient and beam group methods come in first and second place, respectively.

In summary, the linear gradient method produces the highest quality solutions on all patient geometries investigated, and requires a very small number of parameters per beam. Therefore, it provides a helpful compromise between solution quality and the size of the search space. However, since the beam group method is generally the second best method for the patient geometries we have tested, and is the most similar to beamlet-based optimization, we are eager to investigate ways of improving this method in future work (such as by varying the positions and shapes of the beam groups on the BEV). Also, we may eventually attempt full beamlet-based optimization as an 'inner loop' problem similar to the approach of Schreibmann *et al* [10, 11]. An earlier version of PARETO did handle fluence optimization inside an inner loop using a simple linear gradient function. With simple fluence parameterizations, however, it is more efficient to handle BAO and FMO as part of a global, monolithic problem.

We have found that solutions with a relatively high mean fluence complexity (as judged by the modulation index and gradient entropy) generally lie in regions of high

PTV conformity on tradeoff surfaces. We have also seen that runs done with more modulated fields often achieve solutions of superior PTV conformity (see for example, Fig. 3.39). Therefore, the main effect of increasing the fluence complexity is to improve PTV conformity (when OAR reduction parameters are employed). However, occasionally OAR sparing is also improved (see Fig. 3.28a).

Finally, while objective function evaluation time is an important consideration, many of our fluence parameterizations have shown comparable run times for the same number of generations (Fig. 3.47). However, we would hesitate to use the isodose projection method in future work unless the speed of the 3D isosurface calculation could be drastically improved.

3.4.6 Investigation into the effect of applying fluence gradients onto the projections of the OARs in the BEV

We have found that multiplying a linear gradient function by the PTV fluence in regions where OAR projections overlap the PTV in each BEV leads to a significant improvement in solution quality for both the lung and cauda equina patients. We have also found that applying two multiplied gradients to the PTV projection helps to achieve superior PTV conformity (see Section 3.3.3.1). However, the improvement appears to be the greatest when the gradients are specifically applied to the OAR projections (e.g. compare Fig. 3.15b to Fig. 3.49b). Therefore, while most of our effort has been focused on increasing the fluence complexity of the PTV parameterizations, in future work we will focus more on OAR modulation.

3.4.7 Evaluation of metrics used for judging fluence complexity

In this work, we have employed three different metrics for assessing fluence complexity. The modulation index is commonly used in fluence analysis [15], quantifying the typical magnitude of the fluence variation between adjacent pixels. Thus, the modulation index shows the largest relative increase for methods where adding more parameters has a significant effect on local pixel variations, such as with the beam group method. The gradient entropy, on the other hand, measures the spread of fluence variations in the map. Thus, it is most sensitive to methods such as the cosine transform method, where increasing the number of parameters causes a large spread in gradient strength as opposed to producing a large number of strong gradients.

The fractal dimension calculation that we have implemented is based on the autocorrelation as a function of distance (lag) between pixels [27]. A high fractal dimension means that the variation in fluence between pixels does not decrease abruptly as a function of distance (causing a small slope on the variogram), such that there is a small difference in the amount of detail visible at different scales. However, if a map contains many regions of constant intensity, the autocorrelation will not decrease very quickly as a function of lag. Thus, the fractal dimension of maps generated by the isodose projection method tends to be higher than for other methods (see Section 3.3.5). This indicates that the fractal dimension may not be the quantity we are interested in when judging fluence complexity. Furthermore, since the beam group and isodose projection methods show the clearest upward trend in fractal dimension with an increase in the number of parameters, we see that the fractal dimension is most useful in distinguishing between moderate and highly modulated fields, a finding consistent with the work of Nauta *et al*. [27]. (Other parameterizations have been shown to offer only slight increases in modulation by adding more parameters, and the fractal dimension appears to be incapable of detecting these). The usefulness of the fractal dimension is also somewhat limited by the profile length. For the data presented in this work, some rows and columns of the fluence images (with shape defined by the PTV mask in the BEV) are as short as only a few pixels, and so in these cases the correlation as a function of distance is not meaningful. Also, with small profiles and simple gradients, the linear range of the autocorrelation function does not appear to be limited to 10% of the profile length as indicated by Murata and Saito [29].

Finally, the metrics may also be directly compared to each other using fluence maps from runs done with various methods and patient geometries. Figure 3.46a confirms that the modulation index and gradient entropy are not very well correlated for the cosine transform and beam group methods (see Sections 3.4.3.2 and 3.4.3.3), whereas they are slightly more correlated for other methods. The fractal dimension is somewhat correlated to the modulation index (and to a much lesser extent, the gradient entropy) for the beam group and isodose projection methods, again supporting the conclusion that the fractal dimension is useful in distinguishing larger differences in fluence complexity (Fig. 3.46b, c).

3.5 Conclusions

We have investigated several fluence parameterizations on three patient geometries. For each parameterization, we have seen that some increase in fluence complexity (judged most reliably by the modulation index) is helpful in improving solution quality on the lung and cauda equina patients. However, for the spine phantom, the main factor responsible for improving solution quality is the optimized BEV margin parameter. For this simple geometry, the basic weight or single linear gradient methods are superior to other fluence parameterizations that generate more modulated fields. Therefore, the best patient geometries for investigating whether a parameterization offers a high degree of fluence complexity are the lung and cauda equina patients, since in these cases the GA has more incentive to produce highly modulated fields. For these patients, the linear gradient and beam group methods are consistently superior. Therefore, our next step for improving PARETO's fluence optimization capabilities is to maximize the potential of the beam group method by optimizing the shapes and positions of the beam groups on each map. For instance, parameters may be defined as a set of (x, y) position values that are triangulated in order to define beam groups of various shapes. Also, since directly optimizing the fluence modulation over the OAR projections in the BEV is very useful, we will develop further methods to do this in future work. The simple and flexible fluence parameterizations investigated here give PARETO the ability to solve the difficult problem of BAO and FMO in a monolithic formulation where beam angles and fluence parameters are treated equally.

3.6 References

[1] D. Craft and T. Bortfeld, “How many plans are needed in an IMRT multi-objective plan database?,” *Phys. Med. Biol.*, 53 (11), pp. 2785-96, 2008.

[2] D. Craft, T. Halabi, and T. Bortfeld, “Exploration of tradeoffs in intensity-modulated radiotherapy.,” *Phys. Med. Biol.*, 50 (24), pp. 5857-68, 2005.

[3] D. Craft, T. Halabi, H. a Shih, and T. Bortfeld, “An approach for practical multiobjective IMRT treatment planning.,” *Int. J. Radiat. Oncol. Biol. Phys.*, 69 (5), pp. 1600-7, 2007.

[4] D. Craft and M. Monz, “Simultaneous navigation of multiple Pareto surfaces, with an application to multicriteria IMRT planning with multiple beam angle configurations,” *Med. Phys.*, 37 (2), p. 736, 2010.

[5] A. L. Hoffmann, A. Y. D. Siem, D. den Hertog, J. H. a M. Kaanders, and H. Huizenga, “Derivative-free generation and interpolation of convex Pareto optimal IMRT plans.,” *Phys. Med. Biol.*, 51 (24), pp. 6349-69, 2006.

[6] T. S. Hong, D. L. Craft, F. Carlsson, and T. R. Bortfeld, “Multicriteria optimization in intensity-modulated radiation therapy treatment planning for locally advanced cancer of the pancreatic head,” *Int. J. Radiat. Oncol. Biol. Phys.*, 72 (4), pp. 1208-14, 2008.

[7] M. Monz, K. H. Küfer, T. R. Bortfeld, and C. Thieke, “Pareto navigation: algorithmic foundation of interactive multi-criteria IMRT planning.,” *Phys. Med. Biol.*, 53 (4), pp. 985-98, 2008.

[8] T. Spalke, D. Craft, and T. Bortfeld, “Analyzing the main trade-offs in multiobjective radiation therapy treatment planning databases.,” *Phys. Med. Biol.*, 54 (12), pp. 3741-54, 2009.

[9] C. Thieke et al., “A new concept for interactive radiotherapy planning with multicriteria optimization: first clinical evaluation.,” *Radiother. Oncol.*, 85 (2), pp. 292-8, 2007.

[10] E. Schreibmann and L. Xing, “Feasibility study of beam orientation class-solutions for prostate IMRT.,” *Med. Phys.*, 31 (10), pp. 2863-2870, 2004.

[11] E. Schreibmann, M. Lahanas, L. Xing, and D. Baltas, “Multiobjective evolutionary optimization of the number of beams, their orientations and weights for intensity-modulated radiation therapy,” *Phys. Med. Biol.*, 49 (5), pp. 747-770, 2004.

[12] A. B. Pugachev, A. L. Boyer, and L. Xing, “Beam orientation optimization in intensity-modulated radiation treatment planning.,” *Med. Phys.*, 27 (6), pp. 1238-1245, 2000.

[13] J. Fiege, B. Mccurdy, P. Potrebko, H. Champion, and A. Cull, “PARETO : A novel evolutionary optimization approach to multiobjective IMRT planning,” *Med. Phys.*, 38 (9), pp. 5217-5229, 2011.

[14] U. Oelfke, S. Nill, and J. J. Wilkens, “Physical Optimization,” in *Image-Guided IMRT*, T. Bortfeld, R. Schmidt-Ullrich, W. Neve, and D. E. Wazer, Eds. Berlin/Heidelberg: Springer-Verlag, 2006.

[15] N. Giorgia, F. Antonella, V. Eugenio, C. Alessandro, A. Filippo, and C. Luca, “What is an acceptably smoothed fluence? Dosimetric and delivery considerations for dynamic sliding window IMRT.,” *Radiat. Oncol.*, 2, p. 42, 2007.

[16] S. Webb, “A simple method to control aspects of fluence modulation in IMRT planning.,” *Phys. Med. Biol.*, 46 (7), pp. N187-195, 2001.

[17] M. Alber and F. Nüsslin, “Intensity modulated photon beams subject to a minimal surface smoothing constraint.,” *Phys. Med. Biol.*, 45 (5), p. N49-N52, 2000.

[18] W. De Neve et al., “Planning and delivering high doses to targets surrounding the spinal cord at the lower neck and upper mediastinal levels: static beam-segmentation technique executed with a multileaf collimator.,” *Radiother. Oncol.*, 40 (3), pp. 271-9, 1996.

[19] L. L. Kestin et al., “Intensity modulation to improve dose uniformity with tangential breast radiotherapy: initial clinical experience.,” *Int. J. Radiat. Oncol. Biol. Phys.*, 48 (5), pp. 1559-68, 2000.

[20] W. De Gersem, F. Claus, C. De Wagter, B. Van Duyse, and W. De Neve, “Leaf position optimization for step-and-shoot IMRT.,” *Int. J. Radiat. Oncol. Biol. Phys.*, 51 (5), pp. 1371-88, 2001.

[21] D. M. Shepard, M. A. Earl, X. A. Li, S. Naqvi, and C. Yu, “Direct aperture optimization: a turnkey solution for step-and-shoot IMRT.,” *Med. Phys.*, 29 (6), pp. 1007-18, 2002.

[22] A. Rogers and J. D. Fiege, “Gravitational lens modeling with genetic algorithms and particle swarm optimizers,” *Astrophys. J.*, 727 (2), pp. 80-98, 2011.

[23] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 3rd ed. 14 (3). University of Michigan: Addison-Wesley, 1992.

[24] J. D. Fiege, Qubist User's Guide: Optimization, Data-Modeling, and Visualization with the Qubist Global Optimization Toolbox for MATLAB. Winnipeg: nQube Technical Computing Corp., 2010.

[25] S. Webb, "Use of a quantitative index of beam modulation to characterize dose conformality: illustration by a comparison of full beamlet IMRT, few-segment IMRT (fsIMRT) and conformal unmodulated radiotherapy.," *Physics in medicine and biology*, 48 (14), pp. 2051-62, 2003.

[26] J. Llacer, T. D. Solberg, and C. Promberger, "Comparative behaviour of the dynamically penalized likelihood algorithm in inverse radiation therapy planning.," *Physics in medicine and biology*, 46 (10), pp. 2637-63, 2001.

[27] M. Nauta, J. E. Villarreal-Barajas, and M. Tambasco, "Fractal analysis for assessing the level of modulation of IMRT fields," *Med. Phys.*, 38 (10), p. 5385, 2011.

[28] M. Bachmaier and M. Backes, "Variogram or semivariogram? Understanding the variances in a variogram," *Precision Agriculture*, 9 (3), pp. 173-175, 2008.

[29] S. Murata and T. Saito, "The variogram method for a fractal model of a rock joint surface," *Geotechnical and Geological Engineering*, 17, pp. 197-210, 1999.

[30] W. De Neve, "Rationale of Intensity Modulated Radiation Therapy: A Clinician's Point of View," in *Image-Guided IMRT*, T. Bortfeld, R. Schmidt-Ullrich, W. Neve, and D. E. Wazer, Eds. Berlin/Heidelberg: Springer-Verlag, 2006, pp. 3-9

APPENDIX A: Parameter Listing

METHOD	PARAMETER	DESCRIPTION	SEARCH RANGE
Basic Weight	w	Constant fluence value applied to PTV mask	$[0, 1]$
Decoupled Multiple Linear Gradients	θ_x, θ_y	Fluence gradient angle in x and y directions	$\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$
	θ_0	Fluence offset	$[-1, 1]$
	w	Weight of combined gradient map	$[0, 1]$
Coupled Multiple Linear Gradients	θ	Fluence gradient angle	$\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$
	g_0	Fluence offset	$[-1, 1]$
	φ	Polar rotation angle	$[0, \pi]$
	w	Weight of combined gradient map	$[0, 1]$
Cosine Transform	$k_{u,v}$	Amplitudes of frequencies lying within a chosen radius from the origin of the map	$\left[-\frac{0.5}{\bar{g}}, \frac{0.5}{\bar{g}}\right]$
Beam Groups	$h_{x,y}$	Central intensities of beam groups	$[-1, 2]$
Isodose Projections	w_i	Weight of region ‘i’ enclosed by isodose contours	$[0, 2]$
OAR Reduction	q	Reduction parameter for a given OAR	$[0, 1]$
BEV PTV Margin	r_{kernel}	Radius of circular kernel. Sign determines operation (erosion/dilation).	$[-11, 11]$

Table 2: Listing of parameter names for various parameterizations.

METHOD	NUMBER OF PTV FLUENCE PARAMETERS		
Basic Weight	$N_{PTV} = 1$	1	
Multiple Linear Gradients	$N_{PTV} = 3N_{grad}$	N_{grad}	N_{PTV}
		1	3
		2	6
		3	9
		4	12
Cosine Transform	$N_{PTV} \sim \frac{1}{4}\pi\rho^2$	ρ	N_{PTV}
		1	3
		2	6
		3	11
		4	17
		8	58
Beam Groups	$N_{PTV} = (N_{coarse})^2$	N_{coarse}	N_{PTV}
		3	9
		5	25
		7	49
		9	81
Isodose Projections	$N_{PTV} = N_{iso}$	N_{iso}	N_{PTV}
		2	2
		4	4
		6	6
		8	8

Table 3: Numbers of PTV fluence parameters used in this work.

APPENDIX B: Scaling Tests

In an effort to see how run time scales with various quantities, several runs were performed on the spine phantom. In Figure B.1a, the phantom size and the maximum number of beams is kept constant, but the number of beam groups is allowed to vary. A least-squares linear fit to the log-log plot reveals that the run time is nearly independent of the number of fluence parameters (beam groups) since the slope of the fit is only 0.08 ± 0.02. This is due to the fact that with the beam group PTV fluence method, the number of fluence parameters only has a small affect on the time required to evaluate the fitness function in the two-dimensional interpolation over the coarse grid of square beam groups. Figure B.1b shows that when the number of fluence parameters and phantom size are kept constant but the maximum number of beams is allowed to vary, the run time scales approximately as the square root of the number of beams *N* ($\text{run time} = N^{0.48 \pm 0.01}$). For each solution, the time required to evaluate the fitness function scales approximately linearly with the number of rays traced and thus the number of beams. However, the number of beams varies for each solution when beam merging is enabled, and thus the run time does not scale linearly with the maximum number of beams set. Figure B.1c shows runs done with a varying bin factor applied to the spine phantom data. The number of voxels in the phantom affects the resolution of the beam's-eye-view plane and thus the number of rays traced. For a simple cube, the total run time is expected to vary approximately linearly with the number of voxels since the number of rays traced is approximately proportional to L^2, where L is the side length, while the number of steps per ray is approximately L. However, for these runs the run time goes as $N^{0.73}$ where *N* is the number of voxels.

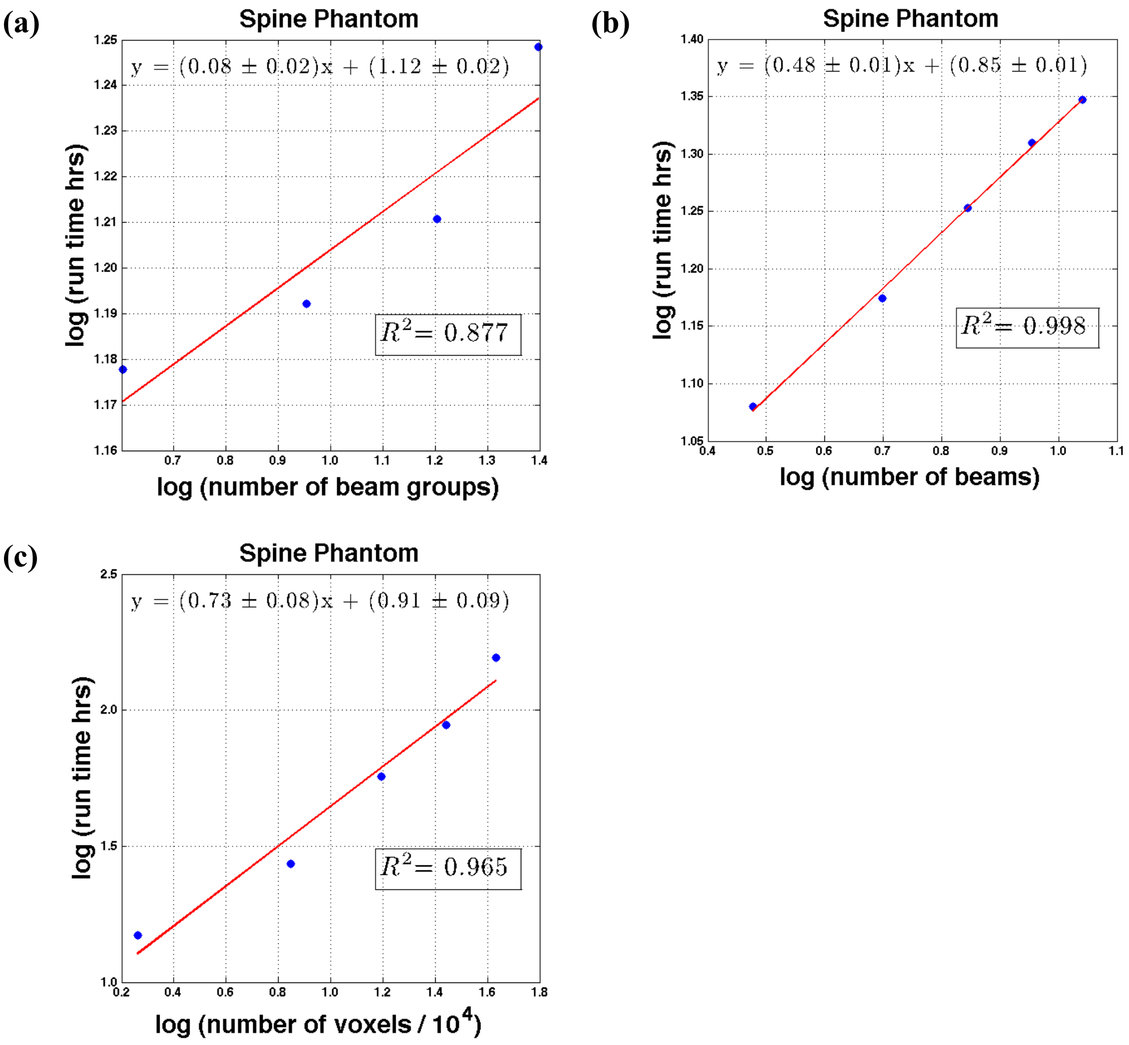

Figure B.1: Run times for spine phantom runs. Red line: a least-squares linear fit with coefficient of determination R^2. (a) Runs with a maximum of 5 beams, 18278 voxels, and a varying number of beam groups for modulation over the PTV projection. Run time is nearly independent of the number of beam groups. (b) Runs with 18278 voxels, one linear gradient for modulation over the PTV projection, and a varying maximum number of beams. Run time is not linear with the number of beams since beam merging was enabled. (c) Runs with a maximum of 5 beams, one linear gradient for modulation over the PTV projection, and a varying number of voxels.