TREATMENT PLANNING ALGORITHMS PART 2 OF 2

Dr Simon Thomas[1] continues his introduction to the main types of photon algorithms used in treatment planning systems, following on from Part 1 which appeared in the previous edition of *Scope*

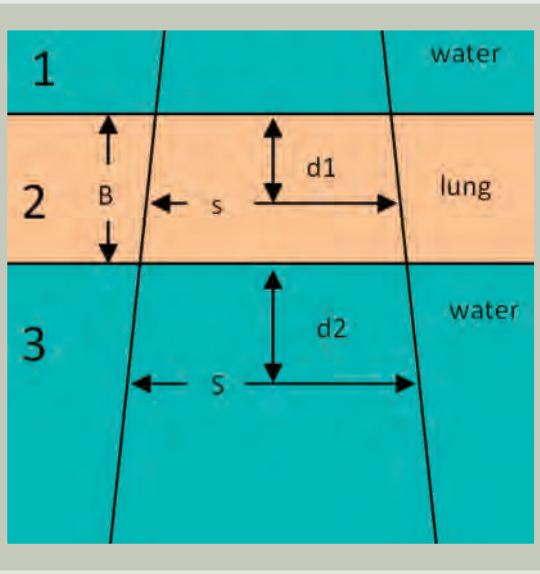

FIGURE 1. Illustration of power law algorithm

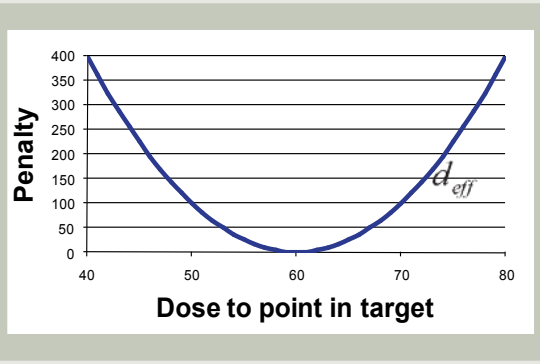

FIGURE 2. Quadratic objective function to a target

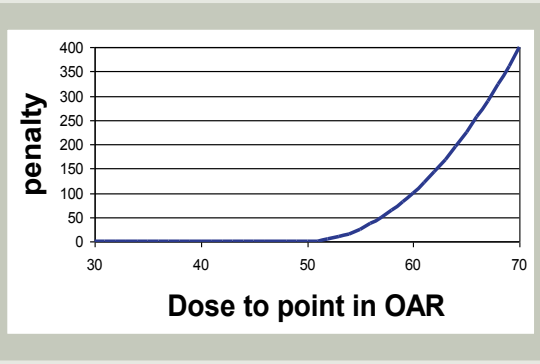

FIGURE 3. Quadratic objective function to an organ at risk

Density correction in 'type a' algorithms

Since patients are not just made of water, some correction needs to be made for inhomogeneities.

Many 1D algorithms have been published; two still in widespread use are the 'effective depth' (otherwise known as 'radiological path length') algorithm and the power law (Modified Batho) algorithm.

Effective depth

The attenuation coefficients for Compton scattering depend on electron density (the number of electrons per unit volume). Electron density is usually quoted relative to water $\rho_{e,w}$.

If you integrate $\rho_{e,w}$ with respect to distance travelled along the ray from the source of radiation to a point of interest, you get an 'effective depth':

$$d_{eff} = \int_{l=0}^{l=d} \rho_{e,w} dl \qquad (1)$$

d_{eff} has many uses in treatment planning, but the simplest is as a means of correcting the dose calculated in water. It can be used to calculate a 'correction factor' (CF) to correct for the fact that patients are not water phantoms:

$$CF = \frac{Dose}{Dose\ if\ water} = \frac{TMR(d_{eff}, S)}{TMR(d, S)}$$

This works equally well with TPR or TAR instead of TMR. It does not work with PDD, since PDD contains the inverse square as well as the attenuation.

Because the dose results from scattered radiation as well as primary radiation, the effective depth algorithm tends to overestimate CF in lungs, especially at the lower end of the MV range (Co-60 to 6 MV). It is for this reason that algorithms such as the power law algorithm were developed.

Modified Batho power law algorithm

The general form of the algorithm takes the form

$$CF = \prod_{m=1}^{m=n} [T(x_m)]^{\frac{\rho_m - \rho_{m-1}}{\rho_0}} \qquad (2)$$

where x_m is the distance transversed in a region of electron density ρ_m,

$$T(d) = TMR(d + d_{max}, S) \text{ and } \rho = \rho_{e,w}$$

To understand how the algorithm works, consider a simple slab phantom as in figure 1. Regions 1 and 3 contain water, whilst region 2 contains lung of density ρ.

In region 2:

$CF = TMR(d1 + d_{max}, S)^{\rho - 1}$

In region 3:

$$CF = \left[\frac{TMR(d2 + d_{max}, S)}{TMR(d2 + B + d_{max}, S)}\right]^{1-\rho}$$

If TMRs were pure exponentials, it can be shown that this would give results identical to the effective depth algorithm. However, since TMR curves fall relatively slowly in the first cm after d_{max}, the Modified Batho algorithm gives slightly smaller CF values than effective depth. For Co-60 to 8 MV this improves agreement with measurement. However, at 15 MV (where scatter is lower, but the TMR curves have a wider peak) Modified Batho under-predicts CF, and effective depth gives better answers.[1]

Inverse planning and optimisation

Objective functions

In traditional 'forward planning' techniques, the user chooses a set of beam parameters (angles, shapes, wedges, etc.), looks at the resulting dose distribution and adjusts the parameters until a good plan is

achieved. In IMRT, the large number of parameters makes this approach impracticable, so instead 'inverse planning' is performed, in which the computer adjusts the parameters until the 'optimal' plan is achieved. For this to happen, one needs a mathematical representation of how good a plan is. An objective function (OF) is essentially just a mathematical measure of how 'good' a plan is. An OF of zero represents the plan that meets all your requirements; the function gets larger the further the plan is from the ideal. There are a large range of functions in use in commercial planning systems. The aim of this tutorial is to give an overview of the types of functions in use. The terms 'objective function', 'cost function' and 'penalty' will be treated as equivalent.

Dose based

The simplest form of objective is to specify a desired uniform dose to a target volume. For example, in figure 2 we are aiming for a dose of 60 Gy. Any voxel in the target that has a dose that is higher or lower than this will add to the objective function, with the penalty typically being proportional to the square of the difference from the required dose.

A similar objective function can be used for organs at risk (*OAR*s), but with the difference that we are specifying a maximum dose to the organ, and have no objection to doses below this dose. We therefore have a one-sided quadratic (figure 3). A shortcoming of this function is that doses only just above the maximum (50 Gy in this example) have only a small penalty, which is why it is often necessary to manipulate the planning system by specifying a maximum that is lower than you will truly accept.

These give the simplest form of a quadratic objective function, summed for all the points in the relevant volume:

$$OF = w_{target} \sum_{n=1}^{N_{target}} (D - D_1)^2 + w_{OAR} \sum_{n=1}^{N_{OAR}} (D - D_2)^2 \cdot H(D - D_2)$$

where w_{target} and w_{OAR} are arbitrary 'weights' or 'importances' (different planning systems use different names) for the target and OAR, respectively. *H* is the Heaviside function, such that:

$H(x) = 1$ for $x \geq 0$, $H(x) = 0$ for $x < 0$

The summation is over all the points in the relevant volume (with rules, depending on the planning system, for dealing with overlapping volumes). The higher the weight for an organ, the more important it becomes to meet the objective for the organ. If the OAR is so close to the target that both objectives cannot be met at once, then choosing w_{target} very much greater than w_{OAR} will ensure that the target objective takes priority over the OAR, whilst choosing w_{OAR} very much greater than w_{target} will cause the OAR to be spared, at the price of target coverage. In practice you will have several targets and *OAR*s, each with their own weights.

DVH based

The main alternative to dose-based objectives is to use objectives that are dose volume histogram (DVH) based.

For a target, the objective will be to pass to the right of a DVH point. For an OAR, the objective will be to pass to the left of a DVH point. A high dose maximum in a target behaves like an OAR. Figure 4 shows examples of DVHs that are failing to meet the objectives.

Various mathematic functions can be used to give a penalty for a failing DVH objective. A simple example is to look at the distance from the point to the curve (Δ in the figure), and to have a penalty equal to $w\Delta^2$, where w is a weight.

Radiobiologically based

A third alternative, which is mainly of interest in a research setting, is to create objectives based on equivalent uniform dose (EUD), tumour control probability (TCP) and normal tissue complication probability (NTCP). It is beyond the scope of this tutorial to go into the details of these models. A change of mindset is required in those who view the resultant plan, as a plan with a uniform dose to the PTV may in fact be worse radiobiologically than one that adds hot-spots to parts of the PTV distant from the OAR.

Optimisation algorithms

The 'optimal' plan is the one that has the lowest objective function. This will only be clinically optimal in cases where the objective function matches the true clinical objective, and avoids a number of pitfalls to be discussed later. For now, we will assume the objective function has been well chosen, and we wish to minimise it.

We will have a number of parameters to vary. What these parameters are will depend on what form of IMRT you are planning. This may be:

- an array of fluences in each of several beams;
- a set of MLC openings and monitor units in 'direct machine parameter optimisation';
- MLC parameters as a function of gantry angle in VMAT, or
- a set of leaf opening times at a number of control points in helical tomotherapy.

Whatever the parameters represent physically, they will reduce mathematically to a set of numbers, subject to sets of rules limiting their values (for example, monitor units cannot be negative). The job of an optimisation algorithm is to choose the set of parameters that gives the lowest possible value of the objective function. Generally there will be hundreds or thousands of parameters being varied. In the examples below I will illustrate algorithms with graphs with one or two parameters; the reader will need to use their imagination to extend these to n-dimensional space.

Figure 5 uses a quadratic objective function (OF) for a target and OAR in a three-field plan with wedges, keeping the two wedged laterals constant at 70 per cent weight, and seeing how the OF varies as the anterior weight varies. For a well-behaved OF like this, it is easy to devise an algorithm to find the minimum, at a beam weight of approximately 110 per cent.

Figure 6 shows a contrived OF curve for a more complex objective function. Here, there is a 'local minimum' at a weight of 90 per cent. An algorithm that finds this minimum runs the risk of erroneously reporting this as the true minimum. This is referred to as 'being trapped in a local minimum'.

Figure 7 shows an OF for a four-field prostate plan, where the lateral weights are constant and the posterior and anterior weights are variable. The objective function is based on the dose-volume histograms to two PTVs, the rectum and the femoral heads, plus a dose-based objective to keep other tissues below 107 per cent.

The global minimum is at 22, 95, but there are a number of local minima.

This plot illustrates the complexity of an objective function, even with only two variables. In most cases of IMRT optimisation the number of variables is

→ AFFILIATION

Head of Medical Physics and Clinical Engineering, Addenbrooke's Hospital, Cambridge

simon.thomas@addenbrookes.nhs.uk

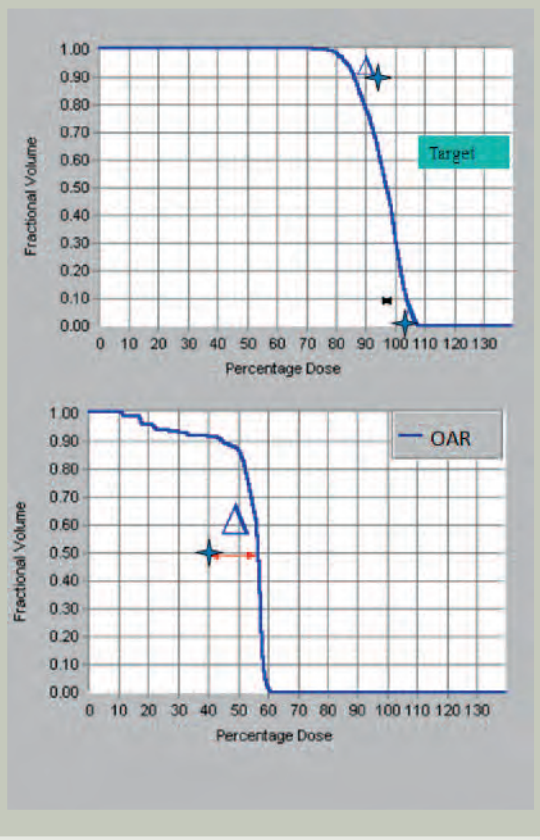

FIGURE 4. The use of DVH as an objective. All objectives in this example are failing to be met

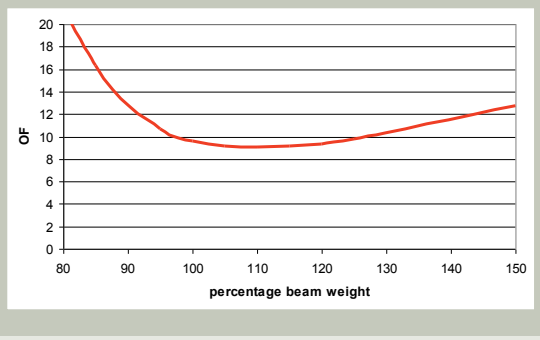

FIGURE 5. A simple OF

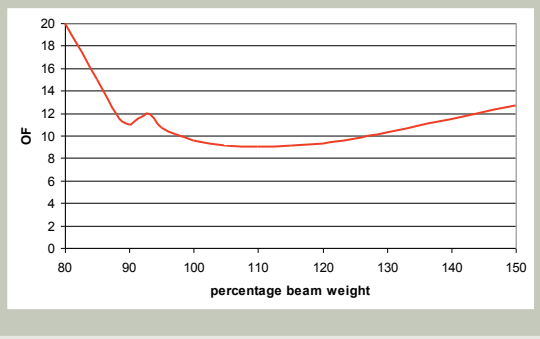

FIGURE 6. OF with local minimum

several thousands, so the topography would have to be seen in n-dimensional space.

The most common approach used in commercial systems is some variety of gradient descent algorithm. In the simple case of a 1D or 2D problem, this can be viewed as analogous to a skier following the slope down to the valley.

For this reason, these algorithms are sometimes known as 'downhill techniques'.

Before considering the multidimensional case, we will consider the maths for a one-dimensional case.

Newton's method

Consider the objective function shown in figure 6, which has a minimum at a beam weight of 107 and a local minimum at a weight of 91.

At the start of optimisation we will have picked an arbitrary beam weight.

Simple down-hill optimisation would keep making small changes in beam weight until it finds a region with zero slope. This would take a large number of steps if the initial value was far from the final solution, and would get trapped in the local minimum for any starting value to the left of the local minimum.

Newton's method calculates the size of the step to move for the next iteration, on the basis of the slope (dy/dx) and the slope of the slope (d^2y/dx^2).

Starting from a beam weight X1, we can estimate the next step X2 using the following:

$$X2 = X1 - (dy/dx)/(d^2y/dx^2)$$

Unless you are unlucky in your choice of starting point, this method will usually jump over the local minimum and converge to a point near the true minimum in only a few iterations.

Newton's method can be extended to the multidimensional case, to calculate how far to go at each step of the iteration. This relies on being able to calculate the first and the second partial derivative with respect to each of the weights.

The matrix containing all the second partial derivatives is known as the Hessian matrix. It can be shown that for a pure quadratic OF, one can get straight to the optimal solution in a single step calculated from the product of the gradient and the inverse of the Hessian.

For more complicated OF, this will usually give a move towards the optimal solution, which is then the basis for the next iteration. There are a number of variants of this, known as 'quasi-Newton' methods, which are beyond the scope of this tutorial. The key points to remember about the method are:

- The objective function must not have any sudden steps in it.
- The slope of the objective function must not have any sudden steps in it.
- Convergence to a solution can be fast.
- The method can get trapped in local minima.

Simulated annealing

A method that was used in some early inverse planning systems (such as Corvus) was simulated annealing. The name is chosen by analogy to the physical process of annealing, which is the name for any technique that involves heating something up and then letting it cool down (figure 8).

When a particle has a high temperature, it can easily jump from one energy state to another. As it cools down, it settles in a single state. Simulated annealing is a mathematical technique that imitates this; initially the parameters can make large changes. Changes that decrease the objective function will always be accepted; at a high temperature some of those that increase it will also be accepted.

As the simulated 'temperature' decreases, the probability of accepting a change that increases the objective function decreases. At low 'temperature', simulated annealing becomes more like a downhill method.

Simulated annealing can deal with objective functions with discontinuities and objective functions with local minima.

Gradient methods (e.g. quasi-Newton) are much faster than simulated annealing. In practice, this means that most commercial planning systems use some sort of gradient method.

In practice, local minima are less of an issue than might be imagined. You should never blindly accept the first 'optimised' solution a planning system comes up with. A small change in objectives and weights, followed by

reoptimising from where you left off, will generally get you out of a local minimum.

Objectives and constraints

I have carefully stuck to the word 'objective' so far in this tutorial, and avoided the word 'constraint'. Whilst some people (and some authors of planning systems) use these terms interchangeably, there is an important distinction that should be made.

An objective is something you would like to achieve, but will compromise on if necessary where it conflicts with other objectives. A constraint is a rule that cannot be broken. Some constraints are physical, such as limitations on MLC leaf speeds, or the inability of linacs to deliver negative radiation (thus setting a constraint that segments cannot have negative monitor units). Constraints can also be dosimetric parameters that will cause a plan to be rejected. For example, an oncologist may wish to keep the dose to the spinal cord below 45 Gy (objective), and will refuse to accept any plan in which the spinal cord exceeds 50 Gy (constraint). This distinction is a useful one when getting oncologists to communicate what they want, and can be used on prescription forms.

However, not all planning systems follow this convention. Some (such as Pinnacle) do, but others use terms such as 'hard' or 'soft' objectives. Some (such as Tomotherapy) call everything a constraint. In practice, in most planning systems, the only way to distinguish them is to give much larger weightings to the objectives that you deem to be constraints.

Pareto optimisation

IMRT optimisation usually has many objectives, some contradictory. For organs at risk near a target volume, there is often a trade-off between different objectives.

The standard solution is to play with the weights of these objectives until you get a plan you are happy with. This involves trial and error by the planner, and the compromise chosen by the planner may not be the same as the one that would have been preferred by the clinician.

A technique that addresses this limitation is known as Pareto optimisation, sometimes referred to as 'multicriterion optimisation'. This looks at how the trade-off occurs, and explores the set of possible plans. The method derives from the work of Vilfredo Pareto (1848–1923), an Italian economist.

For example, consider a plan in which you had set achieving the prescription to the target as a constraint, whilst wishing to keep the doses to the cord and to the lungs as low as possible.

Instead of giving weights to each of these objectives, the system would produce a set of plans, all of which met the target volume constraints, and evaluate the chosen parameter to each organ (which may be mean dose, near-max dose or some biological function such as EUD or NTCP). The set of these would be explored to determine the Pareto front, which is the red line in figure 9.

The planning system would then allow the planner or clinician to scroll through the set of plans, and choose their preferred one. In a real case there will be more than two variables. Interactive displays that allow plans to be selected by multiple sliders have been developed, including the ability to 'lock' an organ.

Pitfalls of inverse planning

Baggy plans that meet all objectives

If you produced a plan optimised with an objective function based on the PTV and all relevant *OAR*s, it would probably not conform tightly to the PTV, since a plan with over-generous coverage will usually result in a slightly more uniform dose to the PTV. One potential way to avoid this is to treat all voxels that are not in any other volume as a remaining volume at risk (RVR) as defined in ICRU83.[2] However, this is not generally enough to ensure that plans conform tightly. In practice, it is necessary to create one or more 'ring' structures around the PTV, setting these as *OAR*s with limits on the dose allowed in these rings.

Fluence loading at skin surface

A common situation in radiotherapy is for the CTV to end a number of millimetres from the skin surface, but for the PTV to extend to the skin; ➤

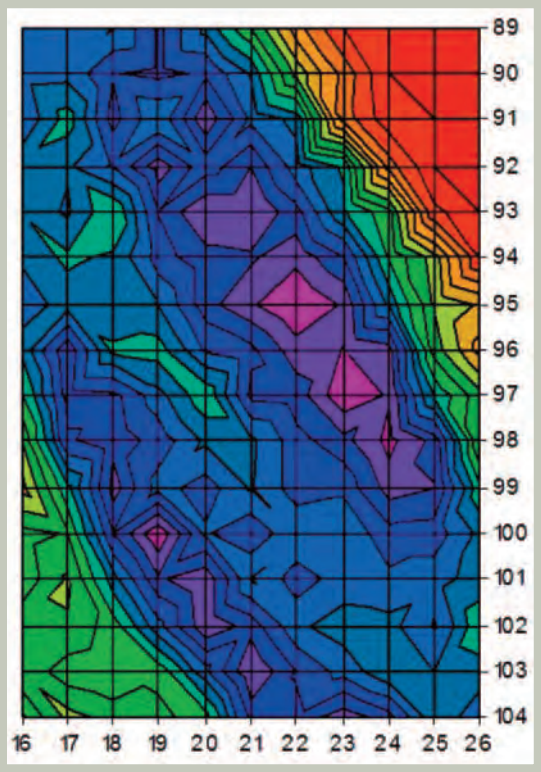

FIGURE 7. 2D OF topography. Red represents the highest OF values, purple/pink the lowest

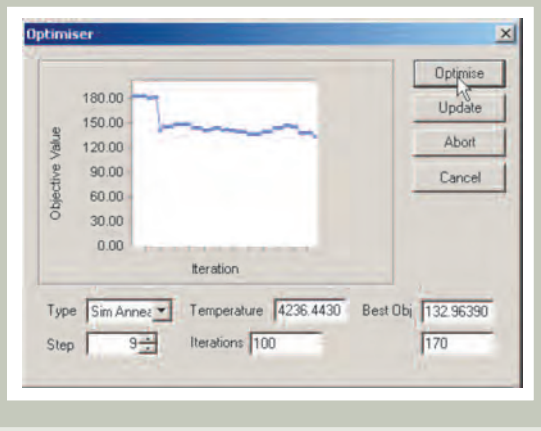

FIGURE 8. Simulated annealing

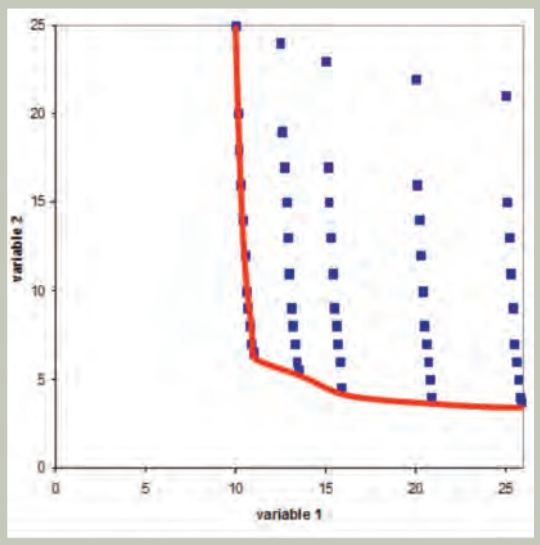

FIGURE 9. The red line represents a Pareto front

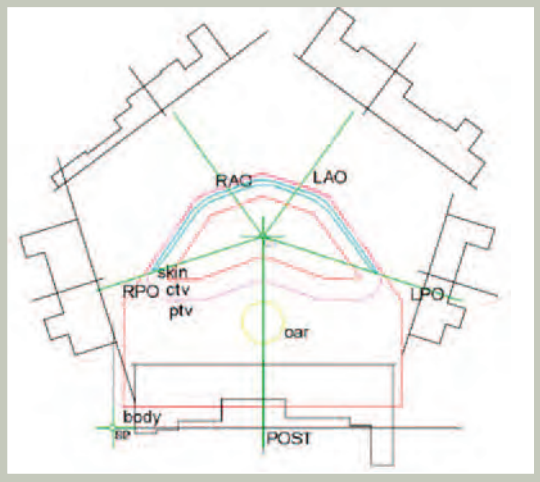

FIGURE 10. The edges of the beams have increased fluence to compensate for build-up

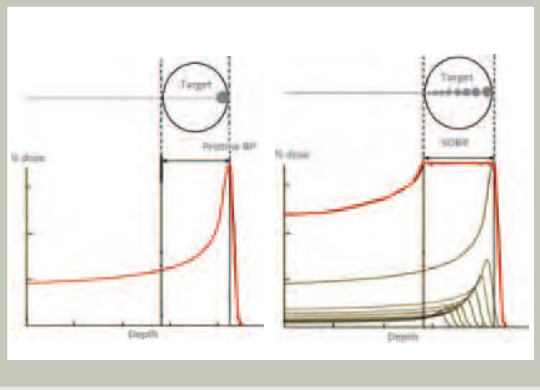

FIGURE 11. A single Bragg Peak and a Spread Out Bragg Peak composed of peaks of multiple energies

there is no disease in the skin, and the oncologist wishes to spare the skin. The PTV margin allows for geometrical uncertainties in setup, but these uncertainties generally preserve the distance of the CTV from the skin.

In a forward planned solution, one ends up with plans where the build-up effect inherently produces the desired skin-sparing. However, in inverse planning, this is seen as an underdose to be addressed, and the algorithm will add fluence into the part of the PTV outside the CTV, as shown in figure 10.[3] This gives unnecessary dose to the skin.

There are two ways to avoid it. One method is to pull the PTV back from the skin. However, this will then mean that the plan is not robust to geometric uncertainties so should only be done if your planning system allows the beams to be extended after optimisation, with a function such as 'add flash'. In planning systems that do not permit the addition of flash, another solution is the use of 'pretend bolus',[3,4] in which the plan contains bolus (not present for treatment) to convince the optimiser that there is no under-dose to correct.

Magnetic fields

The introduction of machines which combine an MRI machine with a linac or cobalt unit means that the effect of magnetic fields needs to be considered. An electron moving through an electric field experiences a force that is perpendicular both to the motion and the field. As a result, kernels calculated in an assumption of no external field cannot be used directly.

Monte Carlo-based systems can model this effect by adding the effect of the magnetic field. The planning system sold with the ViewRay MRI/cobalt system (MRIdian) is Monte Carlo based, with the magnetic field modelled in the calculations. Several groups are developing Monte Carlo algorithms for MRI-linacs; the GPUMCD code from Utrecht is being incorporated into Monaco.[5]

LBTE-based algorithms can in principle be extended to include magnetic fields, although no commercial system currently uses this approach.

Dose calculation for charged particles

The emphasis of this tutorial has been on radiotherapy delivered with x-rays, since this constitutes the majority of radiotherapy delivered in the UK and elsewhere. I will briefly touch on the issues of treatment planning for electrons and for protons.

Electrons

Photon algorithms tend to be good at modelling primary radiation and OK at modelling scatter. Electrons are nearly all scatter, so are much harder to plan accurately. Measured beam algorithms are dangerously wrong in anything but a water phantom. Whilst pencil beam algorithms are OK to indicate qualitative shapes of dose distributions, they have severe limitations when used to make quantitative predictions. The only reliable method for electron planning is Monte Carlo.

The most widely used pencil beam model is the Hogstrom model[6] based on Fermi-Eyges scattering theory of multiple Coulomb scattering. The Gaussian sigma depends on depth (z) and density. Depth dose is based on measured depth dose in water, scaled for effective depth.

This algorithm can work in some cases, but can disagree greatly with measurement, especially at different SSDs from those measured. It tends to underestimate the effects of sharp discontinuities in density, but is useful to produce qualitative plans showing clinicians how much worse the dose distribution is than they imagined.

The discussion of Monte Carlo algorithms in the first part of the tutorial applies also to electrons. VMC++ and MMC are in use in different commercial systems for electron planning. To commission an MC-based electron algorithm, you need accurate knowledge of the beam parameters. These include primary electron energy (0.2 Mev changes range by 1 mm), materials and dimensions of scattering foil, monitor chamber, applicators and cut-outs. You also need accurate knowledge of CT to electron density conversion. It is vital to validate any model against the full range of field sizes and SSDs, and in a variety of heterogeneous phantoms (modelling lung, air, ribs, etc.).

Protons

In contrast to the situation for electrons, 'pencil beam' algorithms give reliable answers in simple phantoms for proton beams. This is because the mass of a proton is 1836 times greater than that of an electron. Hence, the assumption can be made that a proton will usually continue in a straight line as it loses energy by ionisation.

Several manufacturers, including Varian (Eclipse), Elekta (Xio) RaySearch (RayStation) and Philips (Pinnacle) have systems based on pencil beam algorithms. The method of calculation will depend on whether the delivery system uses a passively scattered beam or spot scanning. The depth dose for an individual spot (which follows a Bragg peak such as in figure 11) and the variation of lateral spread with depth can be derived from Monte Carlo calculations, then combined in a series of 2D convolutions in a similar manner to photon pencil beam calculations. The lateral spread results from a number of interactions, including multiple Coulomb scattering and nuclear interactions.[7]

The reason for choosing pencil beams rather than Monte Carlo has been to ensure fast calculations. However, algorithms such as Accuros PT (recently announced by Varian) and VMCpro are claiming acceptable calculation speeds for MC-based solutions.

In x-ray radiotherapy, a change of 1 cm in the effective depth to the target (caused, for example, by the patient gaining or losing weight, or changes in bladder or rectal filling) will cause the dose to change by a few per cent. In contrast for protons, the same change will cause the Bragg peak to shift by 1 cm.

As a result, one may not deliver the plan one was expecting. Since different shifts will come from different beam directions,[8] some plans will be more robust to uncertainties than others. The use of a PTV margin is not necessarily the best solution in highly modulated IMPT plans,[9] and forward planned solutions may give more robust solutions than some inverse planned solutions.

Summary

- Photons: there are a wide range of algorithms as detailed in table 1. Some are faster than others; some are more accurate than others. Some are better in simple geometry; some are better in complex geometry. Always test your system to find its limitations.
- Electrons: with the exception of Monte Carlo, don't believe the answers. Even with MC test your system to find its limitations.
- Protons: several new systems are being developed. Not all plans that look good are robust to geometric uncertainties.
- Read the documentation that comes with your planning system. It will usually explain the algorithms in great detail. An understanding of the algorithm helps to know what to test.

→ **ACKNOWLEDGEMENTS**

I am grateful to my colleagues Chris Rose, Harpreet Dhiraj and Stacey Holloway for comments on versions of this tutorial.

Figure 7 is by David Eaton and figure 11 is by Stacey Holloway.

I am grateful to the vendors for sending me details of how their algorithms have been implemented.

TABLE 1. Types of algorithm for MV x-rays

Does the algorithm model the variation of penumbra with density?	No ('type a')	Yes ('type b')
Not kernel based	Measured beam data models (still used in some independent checking systems and some in-house planning systems)	Monte Carlo: Monaco, iPlan, MultiPlan, MRIdian Linear Boltzman transport equation: Eclipse (Accuros)
Point kernels	FFT convolution: Xio	Collapsed cone superposition: Xio, Pinnacle, RayStation, TomoTherapy, Monaco, OMP
Pencil kernels	'Pencil beams': Monaco, RayStation, Eclipse, iPlan, MultiPlan	AAA: Eclipse

REFERENCES

1 **Thomas SJ.** A modified power-law formula for inhomogeneity corrections in beams of high-energy x rays. *Med Phys* 1991; 18(4): 719.

2 **ICRU.** Prescribing, recording and reporting photon beam IMRT. *J ICRU* 2010; 10(1).

3 **Thomas SJ, Hoole ACF.** The effect of optimization on surface dose in intensity modulated radiotherapy (IMRT). *Phys Med Biol* 2004; 49(21): 4919–28.

4 **Ashburner MJ, Tudor S.** The optimization of superficial planning target volumes (PTVs) with helical tomotherapy. *J Appl Clin Med Phys* 2014; 15(6): 4560.

5 **Ahmad SB, Sarfehnia A, Paudel MR *et al.*** Evaluation of a commercial MRI linac based Monte Carlo dose calculation algorithm with geant 4. *Med Phys* 2016; 43(2): 894–907.

6 **Hogstrom KR, Mills MD, Eyer JA *et al.*** Dosimetric evaluation of a pencil-beam algorithm for electrons employing a two-dimensional heterogeneity correction. *Int J Radiat Oncol* 1984; 10(4): 561–9.

7 **Soukup M, Fippel M, Alber M.** A pencil beam algorithm for intensity modulated proton therapy derived from Monte Carlo simulations. *Phys Med Biol* 2005; 50(21): 5089–104.

8 **McGowan SE, Albertini D, Thomas SJ *et al.*** Defining robustness protocols: a method to include and evaluate robustness in clinical plans. *Phys Med Biol* 2015; 60(7): 2671–84.

9 **Albertini F, Hug EB, Lomax AJ.** Is it necessary to plan with safety margins for actively scanned proton therapy? *Phys Med Biol* 2011; 56(14): 4399–413.

→ **FEEDBACK**
Scope welcomes your feedback!
@IPEMScope