Michael Goitein

BIOLOGICAL AND MEDICAL PHYSICS, BIOMEDICAL ENGINEERING

Radiation Oncology: A Physicist's-Eye View

Springer

1. RADIATION IN THE TREATMENT OF CANCER

INTRODUCTION

The prognosis for someone diagnosed with cancer is not as dire as is commonly believed. Many cancers, such as early stage cancer of the larynx, childhood leukemia, and Hodgkin's disease, are highly curable. Unfortunately, others, such as pancreatic cancer, have a poor prognosis. The curability of most cancers lies somewhere between these extremes.*

Early in its development, a malignant tumor is generally well localized. As most cancers develop, they tend to spread to neighboring lymph nodes and, as metastases, to noncontiguous organs. When the disease is localized, a local treatment such as surgical excision or radiation therapy is indicated. When the tumor is inaccessible or is intimately entwined with a vital anatomic structure,

* Much of the material in this chapter is adapted with permission of the American Institute of Physics from the article of the same title which appeared in the September 2002 issue of Physics Today (pp. 34 to 36) by Boyer AL, Goitein M, Lomax AJ and Pedroni ES.

or when regional spread has occurred, surgery may not be a viable option, and radiation therapy will then be the preferred approach. Distant metastases can, for the most part, only be treated through the use of systemic approaches such as chemotherapy, immunotherapy, or, more futuristically, molecular targeting. Combination therapy – the use of two, or even three, of the approaches just described – is commonly undertaken to manage optimally the local and proven or likely systemic components of the disease. An important rationale for improving local therapy is the observation that the longer a patient has a viable malignant tumor, the more likely that a metastatic "break out" of that cancer will occur – which generally badly compromises the outcome of treatment. Thus "local control" of tumors is necessary for achieving long-term survival.

Overly aggressive surgery or very high doses of radiation and/or chemotherapy can eradicate a cancer with high probability – but, at the cost of causing unacceptable morbidity. Thus, *the art of cancer treatment is in finding the right balance between tumor cure and injury to normal tissues*. Much of the motivation for improving the technology of radiation therapy stems from the desire to reduce the probability of morbidity – which in turn may allow higher doses to be delivered to the tumor with an associated increase in tumor control probability.

TYPES OF RADIATION USED IN THE TREATMENT OF CANCER

Research in physics has contributed directly and indirectly to cancer therapy over the past century. Only months after their discovery by Röntgen in 1895, X-rays were employed to treat a patient with breast cancer. At present, the most commonly employed radiotherapy treatment employs a beam of high-energy X-rays (often described as a *photon beam*) generated external to the patient and directed toward the tumor. Machines containing radioactive ^{60}Co sources are also still in active use in many parts of the world.

Other forms of radiation which have been used in radiation therapy are: electron beams; implanted or inserted radioactive sources (γ, β, and even α emitters are used); neutrons; pi-mesons; protons; and heavier charged ions such as ^{12}C and ^{20}Ne. The bulk of the material in this book relates to the use of external beam therapy using photons. External beam therapy with protons is discussed in Chapters 10 and 11.

WHY DOES RADIATION WORK?

Radiation can cause lethal damage to cells, mainly by forming highly reactive radicals in the intracellular material that can chemically break bonds in DNA, causing a cell to lose its ability to reproduce. The higher the dose, the greater the probability of sterilizing cells. Such damage is experienced both by the malignant cells one is trying to eradicate, and by the cells in the healthy tissues that receive radiation even though one would wish to spare them. There are two elements to the strategy for maintaining the functional competence of the irradiated normal tissues and organs.

First, there appears to be a small and favorable difference between the radiation response of normal and malignant cells that allows preservation of the normal cells that permeate the tumor, and of the nearby tissues that are included in the target volume.[1] The reasons for this difference are complex, not fully understood, and even controversial. The difference is probably due less to differences in intrinsic cellular radiosensitivity than to differences in the genetic machinery activated by radiation, in DNA repair kinetics, and in the mechanisms of cell repopulation – and is counterbalanced by tumor-protective factors such as the substantially greater resistance to radiation of cells in regions of low oxygen tension such as are often found in tumors. To further the differential effect, the dose is usually delivered in small daily increments, termed fractions. This strategy is generally thought to improve substantially the therapeutic advantage as compared with radiation delivered in a single application. Consequently, in conventional radiotherapy, about 30 to 40 daily fractions of approximately 2 Gy each are used.[2] These fractions are typically delivered once a day, with a weekend break, so that a course of radiotherapy will typically last from 5 to 8 weeks. Treatment may also be accelerated, for example, by delivering two fractions per day, or by delivering higher doses per fraction with fewer fractions.

---

[1] One generally defines as the target a volume that includes demonstrable disease, possible subclinical extension of that disease (delineating this is one of the radiation oncologist's arts), and a safety margin for organ and patient motion and technical uncertainties. This is termed the planning target volume (PTV) as more fully described in Chapter 3.

[2] There are particular clinical situations, usually involving relatively small target volumes, in which far fewer fractions, sometimes only one, are employed.

The second element of the strategy for minimizing the probability of normal tissue injury involves the reduction of the dose delivered to normal tissues that are spatially separated from the tumor. This involves manipulating various properties of the therapy beams, as will now be discussed.

A SINGLE TREATMENT BEAM

Figure 1.1 shows a modern radiation therapy machine. X-rays are produced when electrons, accelerated in a linear accelerator, strike a thick high atomic number target, with the X-rays being then shaped by a contoured flattening filter that makes uniform the otherwise forward-peaked X-ray flux. The accelerator, beam transport system, and beam-shaping devices (inset) are all mounted on a gantry which can rotate a full 360° around the patient. The patient lies on a couch that can move in all three translational directions and can rotate about a vertical axis passing through the gantry's isocenter. The shaped

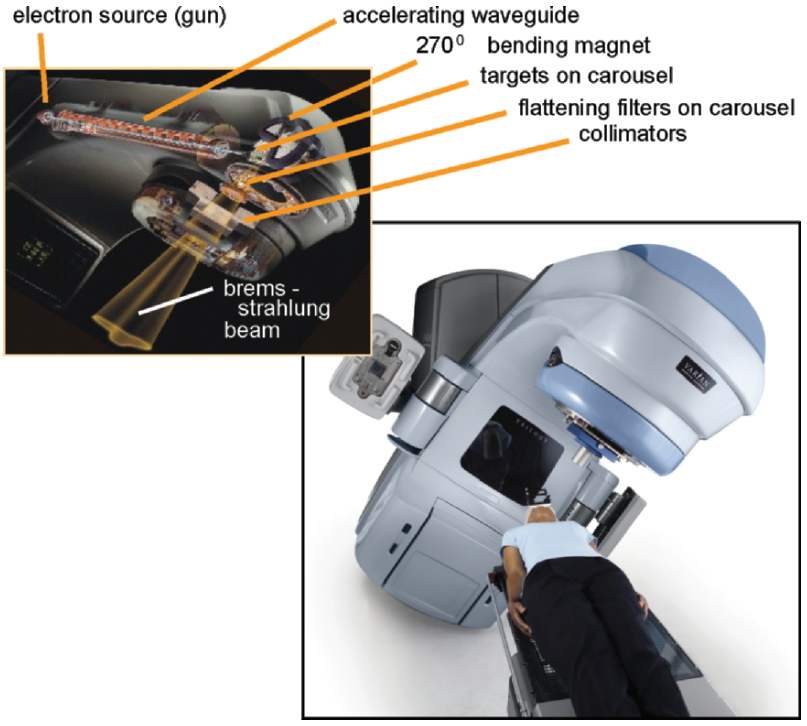


Figure 1.1. A typical modern linear accelerator. Images courtesy of Varian Medical Systems. All rights reserved.

beams of X-rays are directed toward the patient. The beams pass through the patient, undergoing near-exponential attenuation, and deposit dose[3] along the way; the interactions of secondary electrons (see Chapter 4) eventually lead to cell death.

The principal aspects of a single photon beam which can be manipulated are:

- the beam quality (i.e., the maximum photon energy);
- the beam direction (its angle relative to a point within the patient);
- the beam intensity;[4]
- the shape of the field;[5]
- the intensity profile.[6]

How and why these manipulations may be made will be discussed in later chapters. In conventional radiation therapy, as opposed to intensity-modulated radiation therapy which will be discussed below, the intensity profile of the beam is chosen to be as uniform as possible throughout the body of the field, and to fall off as sharply as possible at the field edges. As a consequence, such beams will tend to

---

[3] The dose of radiation is characterized by the energy imparted per unit mass of the irradiated medium. The unit of dose is the Gray (written Gy); 1 Gy = 1 $J \cdot kg^{-1}$ (see Chapter 4).

[4] The term "intensity" is used widely in radiation therapy, but not always in a consistent manner. Indeed, its meaning is context sensitive and often ambiguous. A dictionary definition of "intensity", as used in physics, is "the measurable amount of a property, such as force, brightness, or a magnetic field." (OED, 2001). This leaves open the question of what the property is. And, "intensity" may either refer to the *flux* of the property (e.g., number of photons crossing unit area per unit time), or its *fluence* (e.g., number of photons crossing unit area) which is the integral of flux over time (Webb S and Lomax A, 2001). In talking about dose, the beam intensity may either be understood to refer to the dose rate (dose per unit time), or the total dose. One has to rely on context (and, one hopes, the explicit use of units) to decide which meaning is intended. In the context of the graphic representation of images, "image intensity" usually means the relative fluence of transmitted light through a semi-transparent medium such as film, or of the emitted light from a video display of the image.

[5] The term *field* refers to the area within the lateral margins of the radiation within a plane normal to the beam direction and upstream of the patient.

[6] The term *intensity profile* refers to the lateral distribution of dose within a plane normal to the beam direction.

irradiate the target volume rather uniformly within a plane normal to the beam direction and their dose will diminish near-exponentially along the beam direction.

Multiple Treatment Beams

A single photon beam would, because of the exponential attenuation of X-rays in matter, lead to the delivery of a higher dose to the tissues in front of the tumor than to the tumor itself, as shown in Figure 1.2. In consequence, if one gives a dose sufficient to control the tumor with reasonably high probability, the dose to the upstream tissues would be likely to lead to to unacceptable morbidity. Such a single beam would only be used for very superficial tumors where there is little upstream normal tissue to damage and the skin-sparing properties of X-rays are useful.

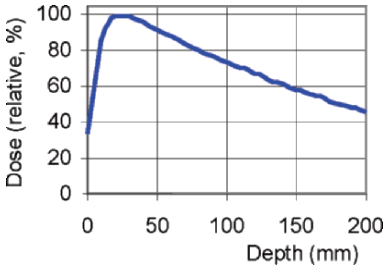


Figure 1.2. Depth–dose curve of a typical 10×10 cm 10 MeV X-ray beam.

The solution is to use multiple cross-firing beams, all focused on and encompassing the tumor, but coming from different directions so as to traverse as far as possible different tissues outside the target volume. This strategy markedly changes the distribution of dose, as is schematically illustrated in Figure 1.3.

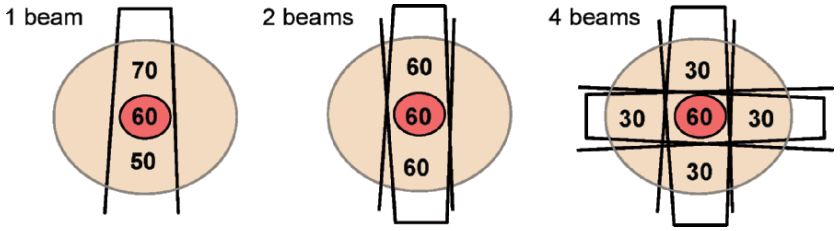


Figure 1.3. Schematic figure showing three different X-ray beam arrangements (1, 2 and 4 beams). Each is designed to deliver 60 Gy at the target center. The (very approximate) dose outside the target is successively reduced as the number of beams increases. However, in a complementary way, the volume outside the target volume that receives an appreciable dose increases as the number of beams increases.

As a result, the dose outside the target volume can, with modern radiotherapy equipment and techniques, often be made to be quite tolerable even at dose levels within the target volume high enough to provide a substantial probability of local tumor control.

THE VOLUME EFFECT

The distribution of dose within the volume of a tumor or a normal organ can dramatically affect its response to radiation. So far as *tumors* are concerned, several points are generally agreed upon:

The larger the tumor, the greater the dose needed to control it. Unfortunately, the larger the tumor the greater the volume of normal tissues which have to be co-irradiated and, consequently, the greater the morbidity. These two effects tend to make larger tumors significantly harder to control than the smaller, earlier stage, tumors.

It is generally desirable to irradiate tumors as uniformly as possible. Dose inhomogeneity is inefficient in that the tumor control probability is likely to be lower than would be the case if the tumor were irradiated uniformly to the same mean dose (Brahme, 1984). In consequence, to achieve the same tumor control probability, the dose would have to be raised – which means that higher doses than necessary would be delivered to the co-irradiated normal tissues.

Nevertheless, some degree of dose inhomogeneity is tolerable. It can even be a good strategy in some situations – for example, in delivering a somewhat reduced dose to the part of a tumor that is closely adjacent to a sensitive normal structure so that the dose received by the sensitive structure can be kept down to a tolerable level, or a somewhat higher dose to a portion of the tumor thought to be particularly radioresistant..

For *normal tissues*, there is no incentive for uniform irradiation – indeed, quite the opposite. For many organs, tolerance is greatly increased when only a part of the organ is included in the high-dose volume – as discussed in Chapter 5. Thus, an important aspect of planning treatments is to take advantage of the volume effect by arranging the treatment beams so that nearby critical organs are only partially irradiated if, as is usually the case, they cannot be entirely spared.

INTENSITY-MODULATED RADIATION THERAPY (IMRT)

So far, we have implicitly assumed that each radiation field is near-uniform over its cross section; dose uniformity of a field within the target volume has, in fact, historically been an explicit goal of radiotherapy[7]. However, the radical suggestion to allow the use of non-uniform fields was made some two decades ago, independently by Anders Brahme and Alan Cormack, fresh from co-inventing the CT scanner – and, in the context of π-meson therapy, Eros Pedroni (Cormack 1987; Brahme 1988; Pedroni 1981). Their idea was based on the judgment that, using mathematical techniques, an irradiation scheme using non-uniform beams could be found which would more closely achieve the ideal of delivering the desired dose to the target volume while limiting the dose to the normal tissues outside the target volume to some predefined value.

Brahme's and Cormack's approaches were motivated by the obsevation that, in CT reconstruction, one can deduce from the intensity reduction of X-rays traversing an object along a series of straight paths what the internal structure of the object is. By inverting the mathematics, one can deduce the intensities (*pencil beam weights*) of a series of very small beams (*pencil beam*s) that pass through the object and deliver dose within it. This procedure leads to highly non-uniform individual fields which, in combination, deliver the desired (usually, uniform) dose to the target volume.

There are two very substantial flaws to the original idea. The first is that, when the problem is posed to deliver zero dose outside the target volume as was initially proposed, many of the computed intensities are negative – a highly unphysical result. The second is that there is no *a priori* way of specifying a physically possible dose distribution to serve as the goal of the optimization.

However, the basic idea of using non-uniform beams has proven enormously fruitful. A workable computational solution is to use optimization algorithms to iteratively adjust the pencil beam weights such that the resulting dose distribution maximizes some score function. The search is computationally intensive and therefore poses interesting technical challenges. However, the still bigger challenge is to find score functions which give a viable measure of clinical

---

[7] Some field "shaping" was used in special circumstances – e.g., wedge filters and, more generally, compensating filters.

goodness. Increasingly, biophysical models of the dose-response of both tumors and normal tissues are being investigated and are beginning to be used as elements of such score functions. These matters are discussed in Chapters 5 and 9.

Intensity-modulated radiation therapy (IMRT), as treatments featuring non-uniform beams are called, has been most intensely developed for X-ray therapy. However, it is equally appropriate for other radiation modalities – including protons. With charged particles one has an extra degree of freedom. One can vary the beam intensity as a function of lateral position *and* as a function of penetration (energy).

TREATMENT DESIGN AND DELIVERY

To design and deliver the best possible therapy, a wide variety of steps have to be undertaken. These include the following.

Tumor and normal tissue delineation

The task of identifying the target(s) and the normal tissues to be avoided or, at least, only moderately irradiated, is described in Chapter 3. Suffice it to say here that it requires:

- the best available imaging to reveal the extent of the tumor and the margins and internal structures of all soft tissues and bones of interest;
- careful control over patient positioning and appreciation of, and sometimes control of, organ motion, as discussed in Chapter 7;
- extensive "manual" effort by experts – only modest progress has been made in automation.

Dose prescription

To plan a course of treatment, clear goals must be set as to how much dose to deliver to the tumor, and how little dose needs to be delivered to the many uninvolved organs and tissues. A so-called *treatment plan* is then developed which defines a set of radiation beams which would deliver the desired dose. The treatment plan, together with a number of other matters such as the fractionation scheme to employ and the way the patient is to be positioned, constitute the prescription.

The prescription is the sole responsibility of the patient's physician. Nevertheless, physicists and dosimetrists have to interact closely with the physician during the development of a plan since the initial prescription goals may be impossible to meet and then some compromise

has to be sought – the nature of which depends on both medical and technical factors.

Treatment planning and evaluation

In modern times, the planning of radiation therapy virtually always takes advantage of computers and of interactive computer graphic displays. The programs that support treatment planning are somewhat analogous to the flight simulators used to train pilots. These programs can simulate:

- the treatment machines available;
- the interactions of radiation with matter;
- the patient geometry;
- treatment delivery of multiple beams from multiple directions;
- the dose distributions resulting from one or more plans.

Overall, the planning process is the task of deciding how to treat a virtual patient with a virtual therapy machine – with the expectation that the simulation is sufficiently good that the actual patient, treated by the actual therapy devices, will receive the desired dose distribution and will experience the best therapeutic result.

The process of treatment planning is discussed in detail in Chapters 6, 8, and 9.

Dose delivery

Once a plan of treatment has been settled on, it must be accurately delivered. This is a complex matter, best summarized by saying that it requires careful attention to a large number of details in order to ensure that the dose distribution that is delivered conforms to that which is desired.

Safety

Above all, a radiation therapy facility must be safe – safer, in fact, than a jumbo jet in terms of the upper limit on the allowable probability of fatalities per mission. All of the steps outlined above, and many more not mentioned, are susceptible to failures in both the procedures and in the underlying hardware and software. A major and critical part of the practice of radiation therapy involves the careful and repeated testing of all parts of the system, and the system as a whole, to ensure that it is as safe as possible.

SUMMARY

The treatment of cancer with radiation can be highly effective. While its effectiveness depends in large part on the type and stage of the tumor and the details of the dose prescription, it also depends on very many technical factors, all of which must be well implemented for a successful outcome. These factors are the focus of the following chapters.

2. UNCERTAINTY

The measured or calculated values of almost all quantities of interest in radiation oncology (and in most other walks of life) cannot be known exactly, but have some degree of uncertainty associated with them. The exceptions are the so-called "denumerable" quantities – things which can be counted such as the number of digits on your right hand. The analysis of the uncertainties in non-denumerable quantities is not an academic exercise, but, rather, is central to the conduct of radiation therapy – not least because it is a life or death matter for patients. As you will see, the following chapters are suffused with talk of uncertainty and the need to estimate it. It is for this reason that I have placed this chapter near the beginning of this book.

While there are very many excellent books on the application of statistical analysis in medicine, I can recommend Mould (1998) as being particularly clear and succinct.

(ALMOST) EVERYTHING IS UNCERTAIN

Each of the following important components of radiation therapy has many aspects with a significant level of uncertainty:

- Diagnosis (e.g., misdiagnosis, wrong histology, wrong staging)
- Imaging (e.g., misinterpretation, spatial distortions, errors in density)
- Delineation of Volumes of Interest (e.g., incorrect tumor identification, incorrect normal tissue identification)
- Prescription (e.g., dosage aims for the target volume and constraints on dose to normal tissues)
- Development of a plan of treatment (e.g., choice of beam direction and field shapes, dose algorithms, plan evaluation)
- Patient handling (e.g., incorrect patient immobilization and/or positioning, errors due to patient and organ motion, changes in patient during therapy)
- Treatment delivery (e.g., incorrect treatment machine configuration, incorrect dose delivery)

One can never ignore the uncertainties; one must deal with them. To do so, one must first have an appreciation of their causes and magnitudes, and of their consequences. Then, to the extent practicable, one should attempt to reduce them to a therapeutically negligible level. If this is not possible, one must adopt some strategy to allow for the residual uncertainties in a manner that is likely to achieve the best result for the patient, such as leaving a safety margin around the tumor volume to allow for motion and alignment uncertainties.

UNCERTAINTY AND ERROR

Physicists are generally comfortable with the proposition that they routinely and unavoidably make errors in their measurements and calculations[1] and are used to analyzing the likely magnitude of the errors and employing tools such as error bars in graphs. Physicians, on the other hand, generally dislike talk of error, perhaps partially for medico-legal reasons, but also for psychological ones. The term "uncertainty" is a bit more comfortable to them and, while I cannot absolve physicians from the need to deal with error, I am happy to be able to reassure them that the word "uncertainty" is the appropriate term to use in characterizing it. For, while the terms "uncertainty"

[1] Uncertainty analysis applies equally to measurements and calculations. In what follows, I use the term "measurement" to refer to both.

and "error" are often used interchangeably, they in fact have somewhat different meanings (ISO, 1995). When one makes a measurement, one virtually always makes an error. That error is unknowable. What one can do is to evaluate in some way the magnitude of the error one is likely to have made and express this as an uncertainty in the measured value. That is, uncertainty expresses the chance that an error of at least a given magnitude has been made. One can have a large uncertainty while having made, in fact, only a very small error.

RANDOM AND SYSTEMATIC ERRORS

When one makes repeated measurements, for example with a ruler of the size of some object, and plots the frequency distribution of the measurements, the result is likely to closely approximate a Gaussian (often termed "normal") distribution – an example of which is shown in Figure 2.2, below. These variations in the results of the measurements are due to *random error*. On the other hand, if the ruler's scale is erroneously calibrated, then, even if the extent of the random errors is negligible, one will consistently and unknowingly make the same error. This is a *systematic error*. Finally, if the observer simply makes a mistake, such as adding a digit to the recorded measurement, he or she has made a *blunder*. The consequences of blunders are usually omitted from uncertainty estimates. The likelihood of many types of blunder can be greatly reduced by double-checking.

It has been traditional to describe the uncertainty associated with these types of errors as random and systematic uncertainties. However, it is now recommended (ISO, 1995) that one refer to type A and type B uncertainties. Rather than focus on the nature of the error, since error is unknowable, the type A and type B designations refer to the way the uncertainty was evaluated. If it was assessed by a "statistical analysis of a series of observations" (e.g., the distribution of the results), then it is termed type A. If by other means, then it is said to be of type B.

In radiation therapy, both random and systematic errors occur. However, as an inspection of the list at the beginning of this chapter will readily suggest, probably the majority of significant errors in radiation therapy are systematic in nature, at least insofar as their repetition throughout a treatment of many fractions is concerned.

PRECISION VS. ACCURACY

Precision and accuracy are often used wrongly and/or interchangeably. They are different. Precision is closely related to random error; accuracy to systematic error. These concepts are illustrated in the target practice scenarios presented in Figure 2.1.

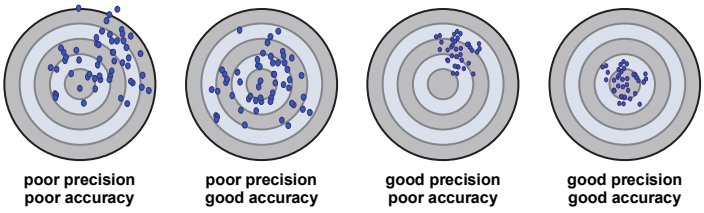


Figure 2.1. Illustration of the concepts of accuracy and presicion.

In radiation therapy, one strives for accuracy as well as precision. It helps the patient little if one consistently gets the same wrong answer or repeatedly does the same wrong thing.

LEVELS OF CONFIDENCE[2]

The shape which characterizes the distribution of uncertainties is called *the probability density function*. In general, the probability density function may have a highly irregular shape. When random errors predominate, the probability density function will be near-Gaussian. When that is so, the shape of the Gaussian function implies that, if the measurement is repeated many times, 68% of those measurements will lie within ±1 standard deviation (often expressed by the letters SD and represented by the greek letter σ) of the mean value – see Figure 2.2.

One common way of expressing the uncertainty in a value is by the size of the standard deviation of its probability density function. Thus for example, if the probability density function can be taken to be a Gaussian function, a value of 2.3 ± 0.2 (SD) indicates that there is judged to be a 68% chance that the true value[3], v, lies in the range of from 2.1 to 2.5.[4,5] The range of values so identified is called the

[2] A further discussion of levels of confidence is presented in Chapter 13.

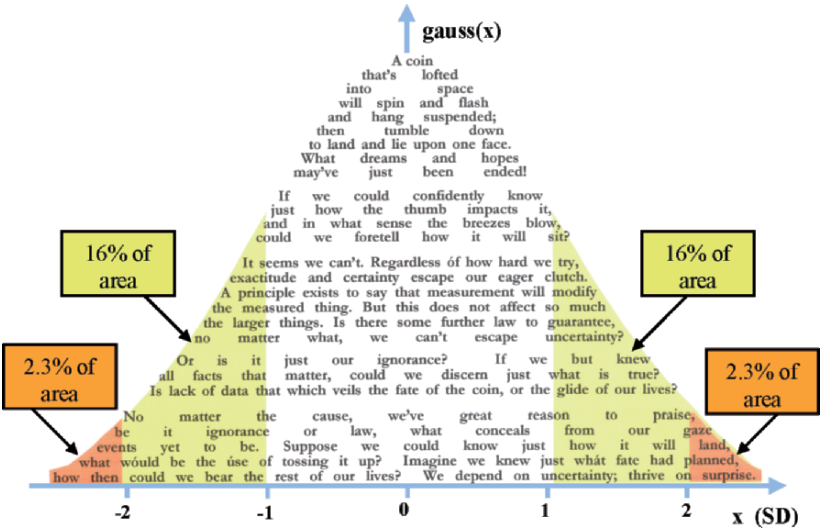


Figure 2.2. The Gaussian distribution. Both the yellow and orange shaded areas extend out to infinity. Verse by the author.

confidence interval. An uncertainty of 1SD is termed a *standard uncertainty*. The uncertainty in a measurement has the same physical units as the measurement itself.

Sometimes it is useful to use the *relative standard uncertainty*, *f*, of a value. This is given by $f = \sigma / v$ and it follows that the standard confidence interval can be expressed as $v \pm \sigma \equiv v\,1 \pm (f)$. The relative uncertainty, being the ratio of two values with the same units, is itself unit-less.

---

[3] On a somewhat pedantic note, the International Organization for Standardization (ISO, 1995) discourages the term "true value" on the grounds that the word "true" is redundant; they take the position that one should not say, for example, "the true value was …"; it is sufficient, and better, to simply state that the "the value was …".

[4] More correctly one should state that, if the value were measured a large number of times, 68% of the time it would fall within ±1SD.

[5] The definition of and estimate of the standard deviation does not depend on the probability density function being Gaussian. One can define a standard deviation for a triangle, a square function, and so forth. However, it is only when the probability density function is Gaussian that the ±1SD interval, for example, corresponds to a 68% level of confidence.

A 68% level of confidence is not a very strong degree of confidence. Another level of confidence which is often used in the medical setting is 95%. A 95% level of confidence means that it is judged that there is only a 5% chance (i.e., one chance in 20) that the true value lies outside the specified range. This corresponds closely, assuming errors to be normally distributed, to 2 standard deviations (1.96 to be more precise) as illustrated in Figure 2.2. One can go further and require a 3 standard deviation interval which corresponds to a 99.7% level of confidence or only one chance in 370 that the true value lies outside the specified interval. However, rarely is one confident enough of the shape of the probability density function to do this.

The 95% level of confidence has received near-religious sanction as the standard required to reach what is called *statistical significance*. Thus, a study comparing a red pill and a blue pill is judged to have shown that they are statistically significantly different in their effects if the difference in their response rate achieves a 95% level of confidence of being nonzero (see Chapter 13).

Representation of the distribution of uncertainties by a single number

The use of a *single number* such as the standard deviation to represent a probability density function is generally a gross simplification – unless one really knows its functional form. In common practice, a Gaussian shape is often assumed and then a single parameter (e.g., its standard deviation) does fully characterize the shape of the entire function.

One-tailed comparisons

Sometimes one is concerned that a measured value not exceed, or not be less than, some specified value. Then, one is only concerned with the likelihood of error on one side of the distribution of errors. As just said, for a normal distribution, a 1SD confidence interval corresponds to a 68% chance that the true value lies inside ±1SD – or, equivalently, a 32% chance that it lies outside that interval. When we only care about values lying to one side of the distribution, this chance is halved. For example, there is only a 16% chance that the true value lies more than 1SD above the measured value (see Figure 2.2).

1.5 standard deviations

A 1 in 3 chance of being wrong is a rather large chance; on the other hand, a 1 in 20 chance is pretty demanding. I have suggested that there is a good case to be made for using a confidence interval which is intermediate between these, namely 1.5 standard deviations, for many situations which crop up in radiation oncology (Goitein, 1983). This is approximately an 85% level of confidence, or a 1 in 7 chance of being wrong. When, as is often the case, one is concerned with a one-tailed issue, then there is only a 1 in 14 chance that the true value lies, say, more than 1SD above the measured value.

Asymmetric uncertainties

The specification of an uncertainty as "±" some value implies that the distribution of errors is expected to be symmetric. While this is often a reasonable assumption, there are circumstances in which symmetric errors are misleadingly wrong. When this is the case, one strategy of desperation is to supply separate positive and negative limits for the uncertainty interval; that is, a probability might be stated as $0.05^{+0.2}_{-0.015}$ (SD). Though not rigorously correct, one could then take the distribution of uncertainties as being represented by two semi-Gaussians on either side of the most probable value, each having a different standard deviation.

Another situation is in the statement of uncertainty associated with a probability. A probability, by definition, must lie between a value of 0 and 1. Thus, a confidence interval on a probability which extends above 1 or below 0 is simply wrong. This gives rise to two problems. First, the uncertainty bounds are certainly usually asymmetric. This is obvious if one considers the case of a measurement which gives a probability of 1 (e.g., 17 of 17 patients responded to therapy). Clearly the upper part of the confidence interval must be zero, whereas the lower part is certainly non-zero. The second problem is that the probability density function must be truncated in order not to have a non-zero value outside the range 0 to 1. It therefore cannot be a Gaussian distribution. The estimation of probability is an example of this. An often-used approach (though, with little justification that I can see other than getting one out of the problem of the limits on the confidence interval) is to assign uncertainty bounds to the logarithm of the probability.

COMBINING UNCERTAINTIES

When analyzing the causes of error in a particular problem, one finds a number of contributing factors; some random, some systematic. In most circumstances, the rule for combining these is very simple:

- make sure that all the uncertainties that are to be combined are associated with the same level of confidence – you don't want to combine standard deviations with 95% confidence limits;
- combine all type A (random) uncertainties in quadrature;[6]
- combine all type B (systematic) uncertainties in quadrature;
- combine the type A and type B uncertainties in quadrature.

– and that is the combined uncertainty, for the same level of confidence as is associated with the individual components. When the individual uncertainties are standard uncertainties, then the combined uncertainty is known as the *combined standard uncertainty*.

In practice, the last three steps of this prescription can be combined. One gets the identical answer if one just combines all uncertainties, of no matter which type, in quadrature. However, knowing the overall type A and type B uncertainties separately can be very informative.

UNCERTAINTY *MUST* BE MADE EXPLICIT

ISO (1995) states that "the result of a measurement [or calculation] … is complete only when accompanied by a statement of uncertainty." Put more strongly, *a measured or computed value which is not accompanied by an uncertainty estimate is meaningless*. One simply does not know what to make of it. For reasons which I do not understand, and vehemently disapprove of, the statement of uncertainty in the clinical setting is very often absent. And, when one is given, it is usually unaccompanied by the qualifying information as to the confidence associated with the stated uncertainty interval – which largely invalidates the statement of uncertainty.

The importance of first estimating and then providing an estimate of uncertainty has led me to promulgate the following law:

---

[6] The sum in quadrature of a set of numbers is the square root of the sum of the squared numbers.

LAW NUMBER 1

When stating a value associated with a measurement or a calculation, one *must*:

(a) provide an estimate of the uncertainty (e.g., through specifying a confidence interval); and

(b) specify the level of confidence associated with that interval (e.g., ±1SD, 95% etc.).

There is simply no excuse for violating either part of Law number 1. The uncertainty estimate may be generic, based on past experience with similar problems; it may be a rough "back-of-the-envelope" calculation; or it may be the result of a detailed analysis of the particular measurement. Sometimes it will be sufficient to provide an umbrella statement such as "all doses have an associated confidence interval of ±2% (SD) unless otherwise noted." In any event, the uncertainty estimate should never be implicit; it should be stated.

In graphical displays such as that of a dose distribution in a two-dimensional plane, the display of uncertainty can be quite challenging. This is for two reasons. First, it imposes an additional dimension of information which must somehow be graphically presented. And second because, in the case, for example, of the value of the dose at a point, the uncertainty may be expressed as either a numerical uncertainty in the dose value, or as a positional uncertainty in terms of the distance of closest approach. One approach to the display of dose uncertainty is shown in Figure 6.4 of Chapter 6.

HOW TO DEAL WITH UNCERTAINTY

To act in the face of uncertainty is to accept risk. Of course, deciding not to act is also an action, and equally involves risk. One's decision as to what action to take, or not to take, should be based on the probability of a given consequence of the action and the importance of that consequence. In medical practice, it is particularly important that the importance assigned to a particular consequence is that of the patient, and not his or her physician. I know a clinician who makes major changes in his therapeutic strategy because of what I consider to be a trivial cosmetic problem. Of course, some patients might not find it trivial at all. So, since he assumes that all patients share his

concern, I judge that he does not reflect the individual patient's opinion very well. Parenthetically, it is impressive how illogically most of us perform our risk analyses, accepting substantial risks such as driving to the airport while refusing other, much smaller ones, such as flying to Paris (Wilson and Crouch, 2001). (I hasten to add that I speak here of the risk of flying, not that of being in Paris.)

People are often puzzled as to how to proceed once they have analyzed and appreciated the full range of factors which make a given value uncertain. How should one act in the face of the uncertainty? Luckily, there is a simple answer to this conundrum, which is tantamount to a tautology. Even though it may be uncertain, *the value that you should use for some quantity as a basis for action is your best estimate of that quantity.* It's as simple as that. You should plunge ahead, using the measured or estimated value as though it were the "truth". There is no more correct approach; one has to act in accordance with the probabilities. To reinforce this point, here is my second law:

LAW NUMBER 2

When faced with uncertainties:

(a) one must assess the odds of the possible outcomes, to the extent feasible;

(b) one must assess the importance of each outcome, be it negative or positive;

(c) then, based on these findings, *one should gamble*.

It may seem irresponsible to promote gambling when there are life-or-death matters for a patient at stake; the word has bad connotations. But in life, since almost everything is uncertain, we in fact gamble all the time. We assess probabilities, take into account the risks, and then act. We have no choice. We could not walk through a doorway if it were otherwise. And that is what we must do in the clinic, too. We cannot be immobilized by uncertainty. We must accept its inevitability and make the best judgment we can, given the state or our knowledge.

3. MAPPING ANATOMY

INTRODUCTION

The possibility of treating deep-seated tumors with radiation depends fundamentally on the ability to "see" – that is, to image – the patient's internal anatomy and, potentially, functional information about it. Otherwise, one would not know what to include and what to exclude from any given radiation beam, or where to aim it. The mapping of

the tumor and normal tissues needs to be done so far as is possible with the patient positioned in a reproducible manner, as discussed in Chapter 7; otherwise the anatomy at the time of treatment may well have shifted relative to where it was at the time of imaging.

Until about the mid-1970s, the principal type of imaging available was X-ray radiography – enhanced by various forms of externally introduced contrast media to image vessels, lymph nodes, body cavities, and so forth. The target volume was typically defined by inference from such radiographs, underpinned by anatomic knowledge and an appreciation of the typical patterns of disease spread. Different parts of the target volume were often defined by different means. Thus, for example, the presence of an abnormally straightened cerebral vessel could indicate the presence of a tumor distending, and hence presumed to abut, the vessel in one region. Elsewhere, the tumor might be presumed to extend up to, but not into, some bony structure thought, on the basis of a bone film, to be uninvolved by disease – and, still elsewhere, by knowledge of a high probability of extension of disease along some anatomic pathway.

Normal (i.e., presumed uninvolved) anatomy was likewise imperfectly determined from radiographs which show bone–tissue and tissue–air interfaces well, but the boundaries between soft tissues at best very poorly. The images were often supplemented by published atlases of normal anatomy derived, for example, by meticulous drawings of cross sections of frozen cadavers, and scaled to match the patient's outer body contour, measured using lead wire.

All this changed dramatically with the clinical availability of computed tomography (CT) – in the mid-1970s for head scans, and in the early 1980s for scanning throughout the body – and, now, magnetic resonance imaging (MRI). Nevertheless, the *principles* for mapping the patient's body remain to this day the same, namely:

- to identify disease where it can be directly imaged;
- to infer the presence or absence of disease from normal tissue abnormalities, or lack thereof;
- to combine information gleaned from multiple imaging techniques;
- to apply knowledge of the known patterns of disease spread; and
- possibly, by marking the anatomy with surgical clips.

The enormous difficulty – indeed, to date, the virtual impossibility – of delineating the target volume by automatic means is due to the

need to combine information from all these elements, using clinical experience and expertise.

Before launching into a description of the various imaging techniques, I want to present the nomenclature which is used for various volumes of interest as some of these are referred to in what follows. The primary goal of imaging is, in fact, to deduce these volumes.

VOLUMES OF INTEREST (GTV, CTV, PTV, OAR ETC.)

The International Commission on Radiation Units and Measurement (ICRU) has done us all a great favor in developing and standardizing a terminology for describing a number of volumes of interest. The ideas were first presented in ICRU50 (1993), and have been clarified and refined in subsequent reports (ICRU62, 1999; ICRU71, 2005; and ICRU78, 2007). I refer you to these publications for the details, and present here a brief summary of the main terms and their acronyms.

Tumor-related terms

Figure 3.1a shows schematically the terms associated with tumor definition. They are:

gross tumor volume	GTV	gross palpable, visible, or clinically demonstrable disease
clinical target volume	CTV	GTV *plus* an extension for subclinical (microscopic) malignant disease
internal target volume	ITV	CTV *plus* an internal margin (IM) for expected physiological movements and temporal variations in size, shape and position of the CTV
planning target volume	PTV	ITV *plus* a setup margin (SM) for uncertainties in patient positioning and alignment of the therapeutic beams

With regard to these definitions:

- Every GTV should have an associated CTV. (However, when it is judged that there is no appreciable extension needed for subclinical disease, the CTV may be identical to the GTV.)
- When there is no gross disease such as when there has been prior surgery with complete resection, only a CTV should be delineated.
- Every CTV should have an associated PTV.
- The delineation of an ITV can be valuable, but is not required.
- The internal margin (IM) and the setup margin (SM) may, and usually should, be added in quadrature, rather than linearly as implied in Figure 3.1a.
- The term "target volume" may be used as a generic term for any of the tumor-related volumes identified above.

I should like to make a few comments regarding these volumes. First, regarding the delineation of the CTV, radiation oncologists often find this task difficult, and it represents a major source of uncertainty. The CTV is certainly not the volume at risk for tumor involvement. That volume is usually the whole body. It is more nearly the volume that has some "reasonable" likelihood of containing tumor and that we think can "reasonably" be treated. The weak point in this is how to define "reasonable". At times the data on which the CTV is based is poor or even anecdotal. The need for clinical judgment in delineating the CTV means that automatic "expansion" of the GTV is generally a poor approach to defining the CTV.

My second point is that, while the GTV and CTV are oncologic concepts, the PTV is a purely physical construction and is intended as a tool to assist in the planning of treatments. It is a limited tool. It does not, for example, tell you where to set the edge of an aperture since it does not include knowledge of the beam penumbra. The aperture must therefore generally be larger than the projected PTV to allow for the beam penumbra. In the case of charged particles such as protons, as discussed in Chapter 11, the PTV may not be very useful in designing beams since the selection of beam range needs a different construct.

Finally, one has to beware that, by defining a sequence of concentric volumes representing a sequence of clinical and physical issues, the margins we allow for these problems tend to be added up linearly, making the treatment fields larger than they need be. Addition in quadrature is usually more appropriate.

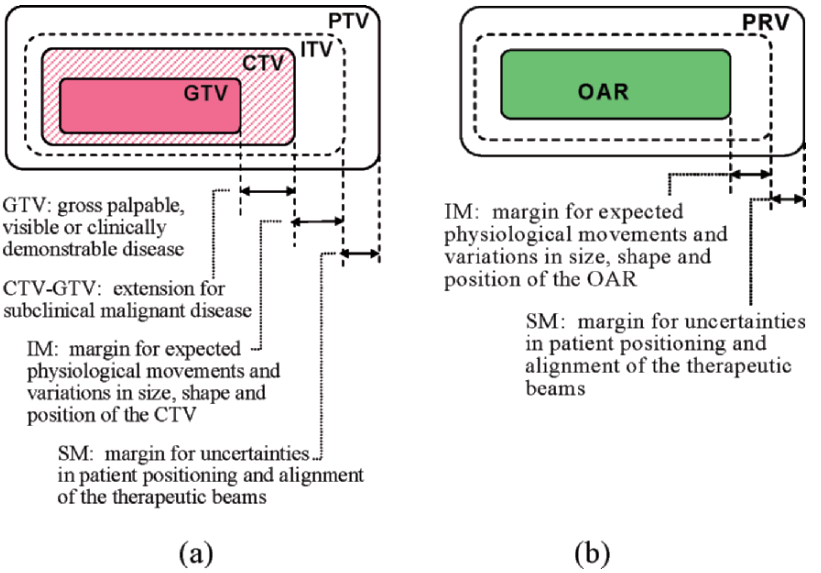


Figure 3.1. ICRU definitions for: (a) tumor-related volumes, and (b) normal tissue-related volumes. Reproduced (in redrawn form) with permission from ICRU78 (2007).

Normal tissue-related terms

Figure 3.1b shows schematically the terms associated with normal tissue definition. They are:

organ at risk	OAR	any organ or compartment of normal tissue which might be significantly impacted by the radiation dose delivered
planning risk volume	PRV	the volume encompassing an OAR *plus* a margin (IM) for expected physiological movements and temporal variations in size, shape and position of the OAR, and *plus* a margin (SM) for uncertainties in patient positioning and alignment of the therapeutic beams

remaining volume at risk	RVR	The volume which is: (a) within the imaged part of the patient, and (b) outside all delineated PRVs (or OARs) and outside the PTV(s)

With regard to these definitions:

- It is desirable that every OAR should have an associated PRV, but it is not required.
- The RVR is always defined, if only implicitly. The reporting of the dose to the RVR, as well as to delineated OARs, is strongly encouraged.

Other terms

Other useful terms are:

Volume of Interest	VOI	any volume which one wishes to define. VOI may be used generically to describe particular volumes such as the PTV, OAR, etc.
Surface of Interest	SOI	the surface of a feature or a plane or curved plane
Point of Interest	POI	any point in space

I strongly encourage you to follow the ICRU terminology. We now turn to the question of how these volumes are determined – or, speaking more generally, to the mapping out of the patient's anatomy.

3D AND 2D IMAGES

The patient is intrinsically multi-dimensional, having multiple properties which are of interest such as X-ray absorption coefficients, T1 and T2 magnetic resonance (MR) decay times and so forth. These properties are distributed throughout the three spatial dimensions and may vary in time The display of any particular property is therefore at least three-dimensional and potentially four-dimensional. At any given time, three spatial dimensions suffice. Thus, to fully represent a region of interest within a patient, one needs to have anatomic information in at least three dimensions. This information may be of many forms. Typically, in CT and MRI for example, some property

of the tissues is measured, resulting in an array of values of that property. From such an array one can generate an image, the intensity of which at any point is proportional to, or related to, the value of the property at that point. The terms *series* and *study* are used here interchangeably to describe a sequence of 2D sections which, together, comprise a 3D data set.

Although the patient extends in three-dimensions, patient anatomy can only be displayed on a screen or on paper as two-dimensional (2D) images. These can be of two basic types:

Sectional images

A sectional image is a two-dimensional representation of a thin "section" or "slice" taken through a 3D data set. Usually the slice is a plane, but it can also be curved. A conventional single-slice CT scanner, for example, produces a sequence of parallel sectional images – usually, but not necessarily, transverse sections. A B-mode ultra-sound scan also provides a sectional image. In sectional images the anatomic information comes from a very thin slice of the patient and, therefore, there is hardly any superimposition of anatomic information; the intensity at a point in the image corresponds to a property of the tissue at that point.

Projection images

The prime example of a projection image is an X-ray radiograph. In such a radiograph, the image intensity at any point in the radiograph is related to the attenuation of the X-rays by all the tissues which lie between the radiation source and the point in question. In this case, the projection is what is termed a perspective projection as the lines appear to diverge from a well-defined point in space. DRRs (see below) are another example of projection images.

An under-valued form of projection image is the photograph. A photograph of the patient's skin surface taken from the presumed source of radiation can be a valuable guide in planning and verifying radiation therapy. A photograph differs from the usual projection image in that it is not so much a superposition of information as a map of the closest visible points on the surface of the patient.

COMPUTED TOMOGRAPHY (CT)

X-radiographs are two-dimensional, formed by superimposing information about the tissues lying between the source and a point in

the image. Thus, three-dimensional information about anatomy is lost. Computed tomography was an enormous technical breakthrough. Using X-ray projection measurements taken from all around the patient's body, CT computes a property of the patient at every point within 3D space to within a spatial resolution of, typically, a millimeter or less. By doing so, CT largely eliminates the confusion of tissues caused by the unavoidable superposition of information in radiographs. The property which is measured is the linear X-ray absorption coefficient of the tissue at a given point relative to the linear X-ray absorption coefficient of water. This is expressed in so-called Hounsfield units (HU), named after one of the inventors of CT. The HU scale is air: -1000, and water: 0. To the extent that the X-ray absorption coefficients of tissues vary from one another, one can then identify the extent of a particular tissue in all three dimensions.[1]

The basis of tomographic reconstruction

How is it possible to reconstruct 3D information from projections? First, CT uses a trick which immediately reduces the problem by one dimension. Namely, it makes its measurements in "slices" through the patient. An X-ray beam is collimated by a slit and the transmitted X-rays measured by a linear detector (for example, an array of small

[1] I made my own entry into medical physics, and became one of the several "inventors" of CT who solved the problem after it had already been solved by Hounsfield and Cormack, through being faced with this problem. While attempting to get a job in Medical Physics I approached Cornelius Tobias at the Lawrence Berkeley Laboratory. He told me of his conviction that one could compute the densities within an object by taking radiographs of it from many different directions. That is, before it was introduced on the public stage, he felt sure that computed tomography was possible. In my haste to impress, I told him that I thought I knew how to solve the problem and would come back in a couple of days to show him the solution. On returning home I realized that I had made a big mistake in my thinking. I had thought that to reconstruct an 80 x 80 map of the object I would have to invert an 80 x 80 matrix – which was time-consuming but do-able. In fact, however, it was a 6400 x 6400 matrix which needed to be inverted, and in those days this was not practical. Pride then required that I find a solution that would work, and I came up with an iterative solution (Goitein, 1972) which I was eventually able to show Dr. Tobias – although he never offered me that job! I offer this anecdote to underline both how varied can be the stimuli to invention and how, quite often, conditions are "ripe" for something to be discovered.

scintillating crystals arranged in a row) as shown in Figure 3.2a. The tube, slit and detector array are then rotated by a small angle and a new measurement is made as depicted in Figure 3.2b – and so on, until the tube has rotated a full 360°. Then, the patient couch may be shifted and a new slice begun.

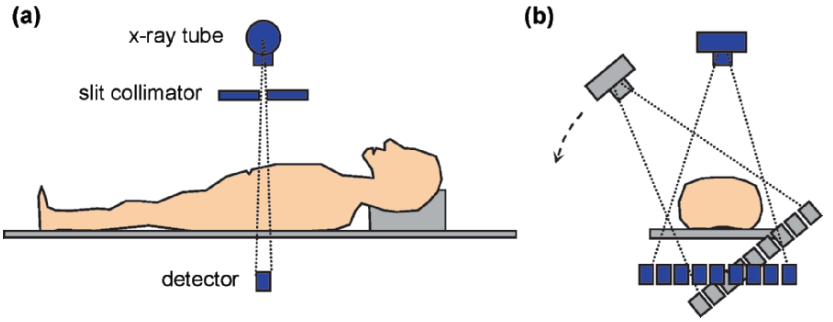


Figure 3.2. Sketch of a one-slice CT scanner. (a) from the side; (b) end-on.

The problem is: how to reconstruct the tissue properties within a 2D slice (averaged over the small slice thickness) from transmission measurements made by a 1D array of hundreds of detectors at a sequence of hundreds of closely spaced angles around the patient? One can see that such a reconstruction is possible from a very simple example which is illustrated in Figure 3.3.

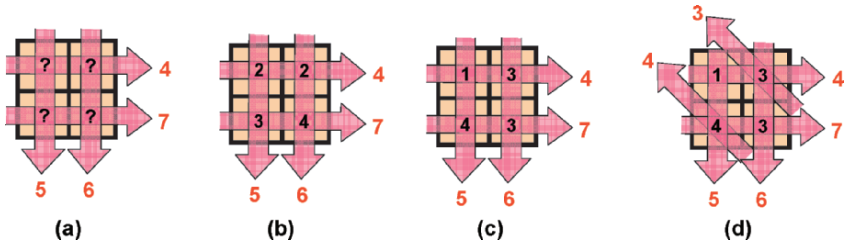


Figure 3.3. CT reconstruction problem in a highly simplified situation (see text). The red numbers are the primary transmission measurements; the the black numbers are the absorptions of each voxel which are deduced from the measurements. All numbers represent percentages (i.e., percentage absorption within voxels; percentage attentuation of measurements).

In this example, a slice of a "patient" is assumed to be made up of only four volume elements (called *voxels*) in a 2×2 array. X-ray transmission measurements are made along parallel paths from two angles (except in the case of Figure 3.3d). The problem is posed in

Figure 3.3a: what X-ray absorption values for each voxel should replace the question marks, given the transmission measurements shown in that Figure?

In Figure 3.3b, values of the X-ray absorption in the four voxels have been estimated by trial and error, and the four values are consistent with the four X-ray transmission measurements. So, have we found the answer? Alas, no – for Figure 3.3c shows 4 different values of X–ray absorption which fit the transmission measurements equally well. What can one do about this ambiguity? The solution is to make additional measurements from one or more additional angles. For example, in Figure 3.3d a pair of measurements at 45° has been added and they unequivocally imply that the solution of Figure 3.3c is the correct one, and not that of Figure 3.3b.

In a general way we can state that one needs redundant measurements – that is, more measurements than unknowns – typically by a factor of three or so. In practice we want to resolve slices with some 512×512 voxels – and so need something of the order of one million measurements. One can readily appreciate that this is computationally highly demanding. The trial and error approach I used to solve the 2×2 problem in my head won't work in the practical situation. Nevertheless, some of the early approaches to CT reconstruction, my own included (Goitein, 1972), used a form of guided trial and error in which an initial guess was refined in an iterative process using, for example, the method of least squares to fit the unknowns (the voxel values) to the measurements (the X-ray transmission along multiple paths). Nowadays a quite different one-pass approach is taken, using Fourier transform methods implemented in special purpose hardware. Also, CT scanners have evolved to, for example, make measurements in multiple contiguous slices and with continuous rotation – sometimes while advancing the patient support continuously (spiral scanning). These improvements have led to much faster scanning and, in combination with respiratory gating for example, allow time-variations of anatomy to be studied.

The information content of CT

In the early days of computed tomography it was hoped that tissues would differ sufficiently in their absorption coefficients that one would be able to identify the histology and pathology of tissues from from their CT values. However, this hope has not been fully realized; while many tissues show marked contrast (bone vs. muscle and muscle vs. fat, for example), most do not. The identification of

specific organs and tissues within a set of CT images is therefore ambiguous and must be supplemented with other information[2]. Nevertheless, the spatial and density resolutions and the speed of CT have evolved enormously so that the current ability to image tissues in space and in time, through the use of gated and repeated studies, is extraordinarily impressive.

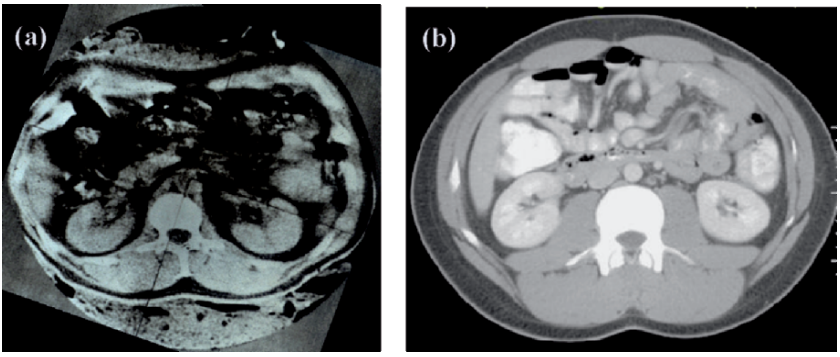


Figure 3.4. Transverse CT scan through the abdomen: (a) one of the very first whole body CT images made public, (b) a scan through a similar body section taken on a modern CT scanner. Panel (b) courtesy of J. Smirniotopoulos, Uniformed Services University, Bethesda, USA.

Figure 3.4a is one of the first whole body CT scans made public, taken on the first EMI whole-body scanner. The field of investigators was small in those days, so I knew the individual whose body was imaged. When in the early 1970s I first saw this image, and several other sections at different levels within the body, I begged copies of them and carried them around in my briefcase for over a year; so impressive to me was the information which they unleashed. Figure 3.4b is a modern CT scan (on an entirely different person) at about the same level of the body, showing how far the technology has come since that first image.

Not only does CT distinguish tissues in space, it does so with very good spatial accuracy. The spatial position of each voxel (volume element) of a CT matrix of values is determined by purely mechanical details of the scanner. As a result, the reconstructed CT values are

[2] Not infrequently, contrast medium is injected intravenously and substantially enhances the contrast between selected tissues.

located at points within a 3D Cartesian grid whose geometry is accurate to within the mechanical accuracy of the scanner, namely at the sub-millimeter level. This spatial accuracy is of particular value in designing radiation treatments as the beams need to be located in space relative to the anatomy to within millimeters and, in certain cases, to within a fraction of a millimeter.

The quantitative information which computed tomography provides is of great use in computing the dose distribution within a patient for a given radiation beam insofar as the influence of tissue heterogeneities is concerned. However, the CT numbers cannot be used directly because they are measured at kilovoltage energies while therapeutic X-rays are in the megavoltage range. As the X-ray absorption coefficient varies with the X-ray energy, the CT values need to be corrected for the energy difference. This is mainly a problem for bone and, secondarily, for fat.[3] The practical solution is to use a look-up table, based on empirical measurements, which converts from Hounsfield numbers (the unit given to the CT values) to megavoltage X-ray absorption coefficients relative to water, as seen in Figure 3.5.

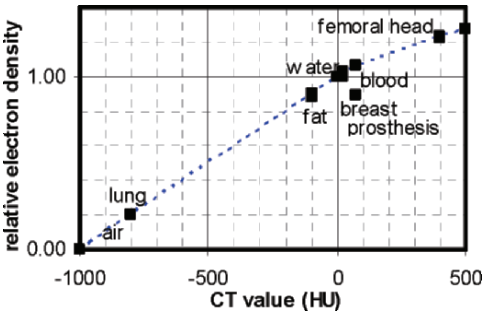


Figure 3.5. CT scanner calibration curve for 135 kVp X-rays. Figure based on data from Battista (1980).

The interpretation of CT images

A discussion of the clinical interpretation of CT or any other images is beyond the scope of this book and beyond the competence of its author. However a few points are in order.

[3] Bone contains a high proportion of calcium which, because of its higher atomic number, causes photoelectric collisions more frequently than do soft tissues. As the probability of a photoelectric collision is highly energy dependent, decreasing roughly as the fourth power of energy as energy increases, the X-ray absorption coefficient of bone is also energy dependent – much more so than soft tissues. Fat is more hydrogenous than most other soft tissues and so, too, has an anomalous effective atomic number.

The conditions under which the images are inspected are critical. The voxels in a typical CT study have values in a range of 1,500 Hounsfield units or more – and differences of about 10 Hounsfield units can be significant. This means that there are at least least 150 different value intervals that need to be distinguished. The values are represented graphically by varying shades of gray in a black-and-white image. However, the eye has a remarkably poor ability, compared to its many other amazing capabilities, to distinguish different levels of grey. Something like 16 different grey levels are all that can be reliably distinguished by most people. This means that one cannot look at a CT image, on either a screen or film, and see all the clinically relevant detail in one picture.

To get around the eye's limitations in this respect, some form of image enhancement is required. By far the most common approach is to process the image such that tissue values below some limit are displayed as black, and above some higher limit are displayed as white, while values in between these limits are assigned intermediate levels of grey according to a linear scale. The central value between the limits is usually termed the "level" and the difference between the limits the "window." Having set a given window and level, one can only distinguish among the limited range of tissues whose Hounsfield units lie within the window. The advantage is that the visibility of, and ability to distinguish between, the tissues within that window is greatly enhanced. The level and window values can be independently varied. Reducing the window size increases contrast – that is, allows a smaller range of values to be inspected with the consequent possibility of greater differentiation among them. On the other hand, adjusting the value of the level while keeping the window constant allows the observer to evaluate a different range of tissue values within the window.

Entirely different information can be obtained from different window and level settings, as Figure 3.6 demonstrates. Window and level settings must be adjusted interactively while viewing CT images. Often, "standard" window and level settings are available – for example, for an overall survey, and to preferentially visualize lung, soft tissue and bone. These standard settings can serve as starting points, but the observer needs to be able to further adjust them.

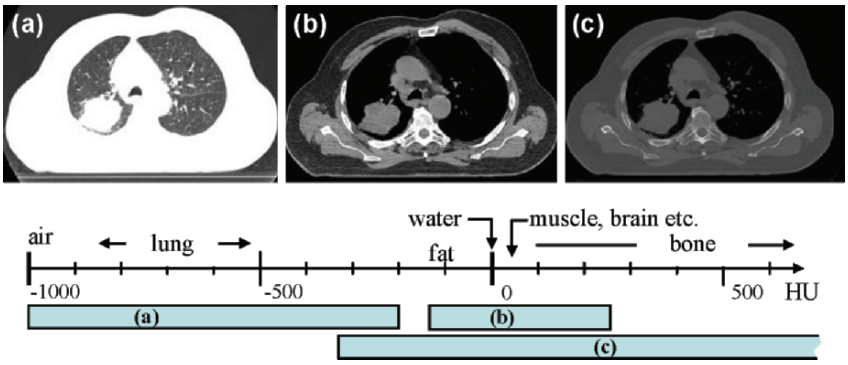


Figure 3.6. CT section through the lung viewed using three different level and window settings: (a) lung settings; (b) tissue settings; (c) bone settings. The lower panel shows the approximate location of tissues on the Hounsfield scale, and the level and window setting used to produce the three images Images courtesy of GTY Chen and M Shinichiro, MGH, USA.

Figure 3.7 shows an example of a CT scan in which the identification of disease in different regions required different window and level settings.

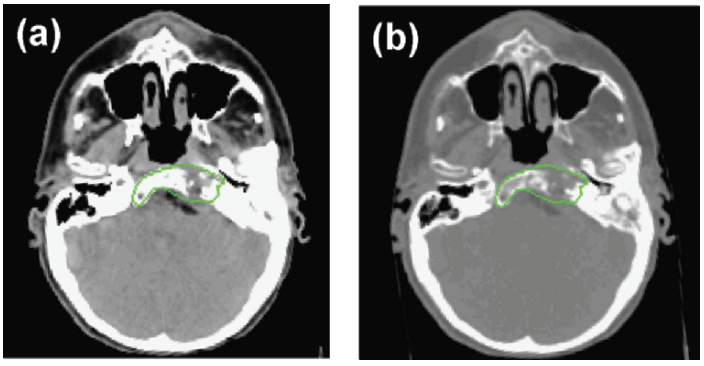


Figure 3.7. Delineation of the GTV required different window and level settings to examine the possible involvement of: (a) the soft tissues (level/ window: 20/135), and (b) the bony anatomy (level/window: 30/670). The The destruction of the clivis by tumor could only be seen clearly in the bone window, whereas the appreciation of the anterior extent of the tumor required the use of soft-tissue settings. Figure courtesy of G. Goitein, PSI, CH.

Re-slicing of the CT Data

CT images are generally obtained as a series of two-dimensional transverse sections[4] in adjacent parallel planes. It is extremely helpful (Goitein and Abrams, 1983) to view the same data resorted into sagittal and coronal planes – and even oblique planes. In early CT studies the spatial resolution along the superior–inferior axis was poor; slice thicknesses and spacings of the order of half to one centimeter were typical. The coarse superior–inferior spatial resolution resulted in "blocky" sagittal and coronal images. This problem has largely disappeared due to scanners with the ability to rapidly accumulate many thin slices with millimeter-scale resolution.

It is particularly helpful to view simultaneously the three orthogonal views; transverse, sagittal and coronal and to identify the planes of intersection on all three images. An example of such a display, including an overlaid dose distribution, is shown in Figure 6.6 of Chapter 6.

Four-dimensional CT (4DCT)

Radiotherapists have known since X-rays were first employed that the patient and his or her internal anatomy are mobile and vary with time. For decades, X-ray fluoroscopy was the best method of evaluating such temporal changes – provided that the region of interest was directly or indirectly identifiable.

In recent years, a major advance in CT imaging has allowed one to synchronize scans with a timing signal (such as a signal at some phase of the respiratory cycle) and to produce CT data sets which show the patient's anatomy at several different times (e.g., at several different phases of the respiratory cycle). Figure 3.8 shows a sequence of such scans for a patient with a lung tumor. This technology allows a quantitative assessment of the degree of motion, and permits tailoring

[4] The principal anatomic planes for imaging purposes are: *transverse* planes, perpendicular to the long axis of the body which divide the body into top (superior or cephelad) and bottom (inferior or caudad) parts; *sagittal* planes, perpendicular to transverse planes which divide the body into left and right parts; and *coronal* planes, perpendicular to both transverse and sagittal planes which divide the body into front (anterior) and back (posterior) parts. Examples of such sections are: transverse – Figure 3.7; sagittal – Figure 3.11; and coronal – Figure 3.8.

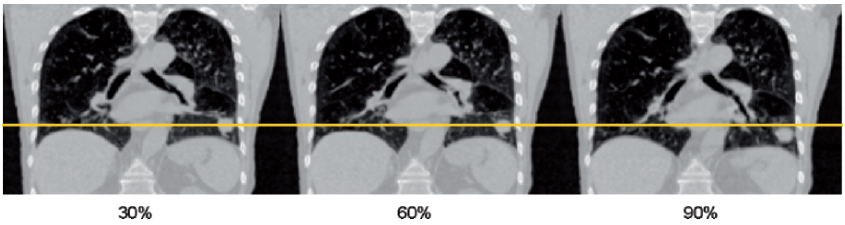


Figure 3.8. Coronal (reconstructed from axial slices) 4D-CT sections of a patient with a small lung tumor. Each image is labeled with the phase of the respiratory cycle. The cephelad-caudad position of the tumor and of the diaphragm changes with phase (e.g., relative to the horizontal line) by about 1.2 cm between the 30% and 90% phases. (Figure courtesy of S. Vedam, MD Anderson Cancer Center, USA).

of the treatment to a specific phase or range of phases of the respiratory cycle. These matters are discussed further in Chapter 7.

Digitally reconstructed radiograph (DRR)

A CT (or MRI) data set is a 3D map of the patient's anatomy. Using this map, it is possible to compute what a radiograph taken from any vantage point would look like. Such a computed radiograph is termed a digitally reconstructed radiograph (DRR). One particularly interesting DRR is that in which the vantage point is the source of a radiation beam; this provides a DRR in the so-called beam's-eye view (BEV). When the treatment collimator and/or aperture are superimposed, such a DRR shows what anatomy is included in, and what excluded from, the treatment beam. Another pair of interesting vantage points are the focal points of X-ray tubes used for alignment purposes. Such DRRs are very helpful for confirming patient position, as discussed in Chapter 7.

A DRR is computed by casting a set of rays diverging from the imagined source of an X-ray tube and passing through the CT data set. The DRR value at the end of any ray is equal to the sum of the CT values (i.e., Hounsfield units, or some property derived from them) along that ray (Goitein *et al.*, 1983). This process is illustrated in Figure 3.9a and an example of such a so-called digitally reconstructed radiograph (DRR) is shown in Figure 3.9b. In summing values along the rays, one can, on the one hand, estimate X-ray absorption coefficients from the Hounsfield units and thereby simulate an actual radiograph, or, on the other hand, one can choose only to sum up Hounsfield units within some limited range (e.g., the

range of values associated with bone) and hence generate an image which provides high contrast for the tissues which lie within that range (e.g., bone). The latter approach was used for the DRR shown in Figure 3.9b.

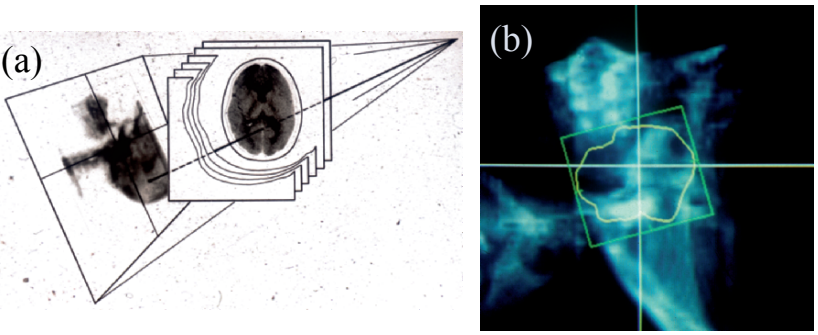


Figure 3.9. (a) illustration of how a DRR is made; (b) the DRR itself with the beam aperture superimposed. Reproduced with permission from Goitein *et al.* (1983).

The DRR shown in Figure 3.9 is an historical (*circa* 1984) image. Much better quality images can be obtained now as a result of better and more finely spaced CT slices and improved algorithms for computing the DRR. Figure 3.10 shows a pair of more modern DRRs.

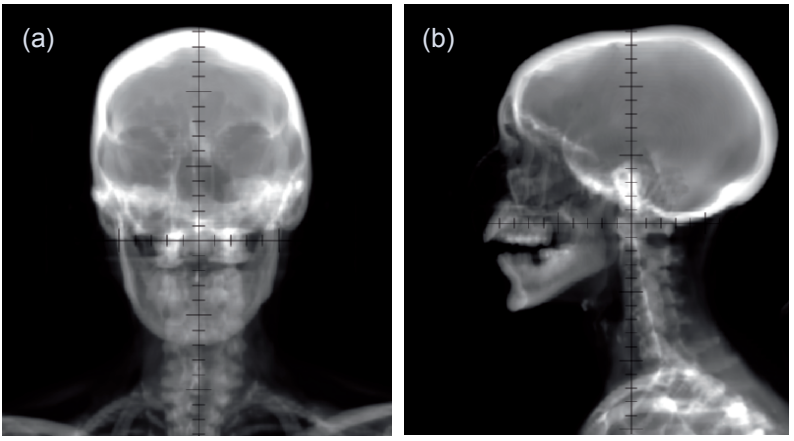


Figure 3.10. High resolution DRRs: (a) anterior, and (b) lateral projection. Images courtesy of L. Dong, MD Anderson, USA.

MAGNETIC RESONANCE IMAGING (MRI)

Principle of operation

This is not the place for a description of what makes magnetic resonance imaging (MRI) possible and how it is performed – nor would I be a good guide to these matters. However, I will make a very few broad comments, principally to underline the ways in which CT and MRI differ from the point of view of their use in radiation oncology.

One of the properties of atomic nuclei is that they act like tiny magnets, their magnetic field arising from the fact that the charged nucleus has a spin and a moving charge creates a magnetic field (which is the basis of operation of electromagnets). The nucleus of interest in MRI is the proton[5] – abundant in tissue not least because it is the nucleus of hydrogen which, in the form of water, constitutes a large fraction of the human body.

A proton's spin, and hence the direction of its magnetic field, is quantized so that it can align in one of two possible directions – crudely speaking "up" and "down." Under normal conditions, the spins are equally distributed between the "up" and "down" directions. However, when a magnetic field is applied to a body of tissue, the proton spins have a slight tendency to align with the magnetic field – thus creating a slight excess of "up" as opposed to "down" protons. The difference in the numbers is very small – typically about one part per million – but is enough to form the basis of MRI. Not only do the spins tend to align with the magnetic field, they also rotate (precess) about it, much in the manner in which a spinning top wobbles (precesses) around the vertical due to the influence of gravity. The frequency of this precession is known as the Larmor frequency and is proportional to the strength of the applied magnetic field. For protons, the Larmor frequency in a 1 T field is 43 MHz. At, say, higher field strengths, the Larmor frequency is proportionately higher.

If a radiofrequency (rf) field of precisely the Larmor frequency is applied to tissue within a magnetic field, the spins tend to be forced into the more energetic down state – and, after the rf field is removed, the spins gradually return to their equilibrium distribution. As they do

---

[5] Except in the realm of magnetic resonance spectroscopy, where other nuclei than protons may be imaged.

so, they emit rf radiation at the same Larmor frequency. I think of it as the spins being excited by the initiating rf field and then giving out little rf "yelps" as they de-excite. The sound of the yelp, that is its frequency, depends on the magnetic field strength; it is the Larmor frequency. If a second receiver rf coil is placed nearby, it can detect the emitted rf radiation and thus "sense" the presence of protons.

If the tissue sample (e.g., a patient) were exposed to a uniform magnetic field, then all the protons within the sample would precess at the same frequency, they would all be excited by the same rf signal, and would all contribute to the received rf signal. That is, there would be no information about the spatial location of the protons. Imagine, on the other hand, that an inhomogeneous magnetic field could be applied in such a way that each small volume of tissue (voxel) experienced a different magnetic field from all others. Then, the Larmor frequency of the protons in a given voxel would be different from that of all other voxels since the magnetic field would be different from that in all other voxels. If an rf field were applied of exactly that voxel's particular Larmor frequency, then only the protons in that voxel would be excited and then the detected rf signal would come from the de-excitation of those protons alone. By sequentially varying the applied rf field in time, one could obtain a set of detected signals, each unique to a single voxel. That is, one would have complete 3D information as to the behavior of the protons in each voxel, independently.

Unfortunately, the laws of physics (in particular, Maxwell's laws of electromagnetism) do not allow one to design a magnetic field whose strength is different at each point in 3D space. But, one can easily achieve partial spatial information by applying a magnetic field gradient across the tissue sample, thus creating planes of different magnetic strength. The received rf signal in that case would come from all the protons in a plane. This is, in fact, the first step in obtaining 3D information in MRI. How the remaining two dimensions of decoding are disentangled is beyond the scope of this short summary. Suffice it to say that, by applying various magnetic field gradients in sequence, and by processing the received signals, one can analyze the signals so as to isolate the rf response of the protons in each voxel of a 3D array of voxels. Since the information is truly three dimensional, it is as easy to display a sagittal or coronal section as it is to display a transverse section.

It is important to appreciate that the spatial information obtained in MRI derives from a knowledge of the strength of the magnetic field at each point in space at any given moment, a point to which we will return shortly. In addition, a wide variety of artifacts can give rise to spatial distortions of the images.

The "meaning" of the received signals depends on: the timing of the exciting rf "pulse"; the timing of the signal received from proton de-excitation; and the timing of the changes in magnetic field. One of the charms of MRI is that, because there are so many variables, there is a wide variety of information that can be obtained. Again, these are too numerous, the process by which they are obtained too complex, and the clinical implications far too broad for me to attempt to describe them here. The three earliest forms of MRI produced so-called: proton density-weighted, T1-weighted, and T2- weighted images. As their names imply, none of these are pure measurements; rather, the specified property is enhanced by technical means.

Proton-density images, as their name implies, give values dominated simply by the density of free protons in each voxel. These images tend to be of rather low contrast. The symbols T1 and T2 refer to different relaxation times – i.e., times for the signal from de-exciting protons to decay away. T2 is the time by which a component of the emitted signal decays away due to so-called spin-spin interactions – that is, interactions between neighboring excited protons. T1 is the time by which a component of the emitted signal decays away due to so-called spin-lattice interactions – that is, interactions between an excited proton and the molecular structure of its environment. Typically, T2 relaxation times are quite a bit shorter than T1 times. Both types of image tend to have good contrast.

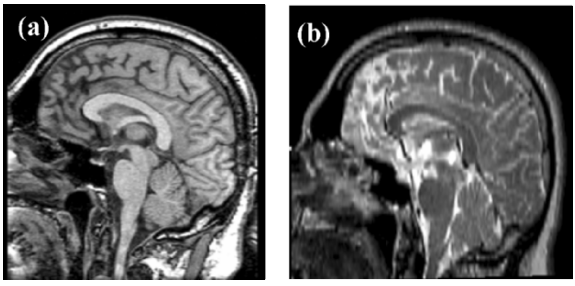


Figure 3.11. Sagittal sections of an MRI study: (a) T1-weighted; (b) T2-weighted. Figure courtesy of the Whole Brain Atlas.

To give a small appreciation of the quality and enormous wealth of information in MRI, Figures 3.11a and 3.11b show typical T1-weighted and T2-weighted sagittal sections through the head; and Figure 3.12 shows a coronal image through the thorax.

A further capability of MRI is termed MR spectroscopy (MRS). The basis for MRS of atomic protons is the fact that the precession frequency of protons is slightly affected by the chemical nature of the proton's environment. Other nuclei and the surrounding electrons slightly modify the magnetic field experienced by a proton, thus changing its Larmor frequency. The consequence is that the emitted signal has a fine structure – that is, a range of frequencies. This spectrum can be analyzed to reveal the nature and concentration of chemicals within a voxel. Perhaps the most common feature to be extracted is the ratio of concentrations of cystine and choline within each voxel.

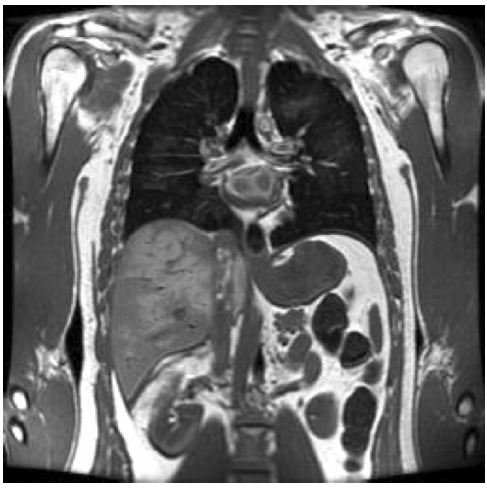


Figure 3.12. Coronal section of an MRI study showing the wealth of detail which can be imaged. Figure courtesy of U.S. National Library of Medicine's Visible Human Project.

COMPARISON OF CT AND MRI

From the point of view of radiation oncology, CT and MRI differ in a few important ways which I will now briefly discuss.

Spatial accuracy

The signal assigned to a given voxel in the reconstructed data set for an MRI study is associated with a particular frequency of the received rf signal. This frequency, in turn, is associated with a particular magnetic field through the Larmor effect. Thus, the spatial location of a given voxel in the reconstructed data set depends on knowing the spatial distribution of magnetic fields. The magnetic fields – the sum of the main magnetic field plus applied magnetic gradients – do not vary perfectly linearly in space and are subject to changes in time for a number of reasons. Thus, MR images are subject to considerably greater spatial uncertainties and non-linearities than are CT images whose spatial accuracy depends only on well-controlled mechanical factors. As a consequence, MR images are generally considered too

inaccurate to use for the accurate geometrical design of radiation beams and, even if the MRI images are clinically superior, a so-called "planning CT" taken with the patient immobilized and positioned as for treatment is generally considered necessary, in addition, for planning radiation therapy.

The visualization of bone in MRI

Compact bone, with its high concentration of calcium, contains very few free protons and, hence, yields very little, if any, MR signal. This lack of signal from compact bone has given rise to the mistaken impression that MRI does not "image" bone. In fact, there is a very high contrast between bone and neighboring tissue; it is just that, while bone appears white against a darker background in CT images, it appears black against a lighter background in MRI images. One has but to reverse the MRI image as in Figure 3.12 to see how strong the bone contrast is. To the extent that bone is eroded by tumor, the signal from the tumor cells will be visible and the bone erosion will be evident.

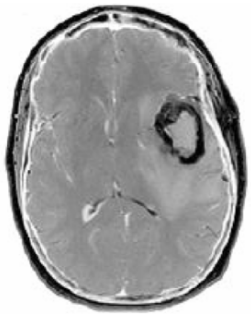


Figure 3.13. MR image with black and white reversed, showing the bone detail inherent in MRI. Original MRI courtesy of M. Kessler.

Clinical information

First and most obviously, as already discussed, CT and MRI scanners measure different properties of matter. Therefore, they provide the clinician with different information concerning the biological properties of the patient's tissues. Whether one or the other provides the more valuable information from the point of view of clinical interpretation depends on the site and the tissues being evaluated and many other factors. However, the two imaging modalities often complement each other, building up a picture of the patient's disease that neither alone can provide. Here, courtesy of N. Liebsch of the Massachusetts General Hospital, are two examples of complementary studies.

Case 1

Figure 3.14 shows two scans of a patient with a low-grade chondrosarcoma. Figure 3.14a is an axial CT section of the patient, injected with intravenous contrast medium. Figure 3.14b is an

MRI (T2-weighted) axial section of the patient at the same level. A low-attenuation tumor is seen in the CT study (see arrow), centered about the left jugular foramen. The MRI study shows extension of the tumor within the bone marrow space of the lower clivus all the way to the contralateral hypoglossal canal (see arrows). This extension was not appreciated in the CT study at any window/level setting and the MRI scan substantially altered the plan of treatment.

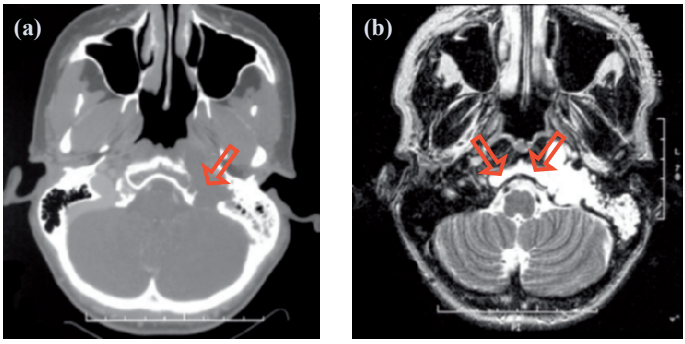


Figure 3.14. (a) Axial CT image with contrast, and (b) axial MR image (T2-weighted) of a patient with a low-grade chondrosarcoma (see text). Figure courtesy of N. Liebsch, MGH, USA.

Case 2

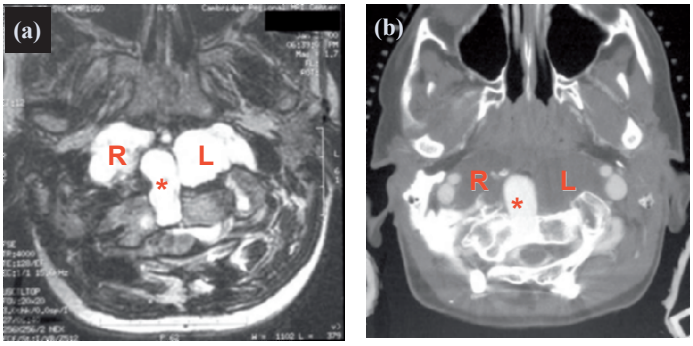


Figure 3.15. (a) Axial MR image (T2-weighted) at the level of C2, and (b) axial CT image with a double-contrast myelogram of a patient at the same level with a chordoma of the cranio-cervical junction after partial resection (see text). Figure courtesy of N. Liebsch, MGH, USA.

Figure 3.15 shows two scans of a patient with a chordoma of the cranio-cervical junction after partial resection, including

odontoidectomy. Figure 3.15a is an axial MRI section (T2-weighted) at the level of the C2 vertebra which shows a multi-lobulated tumor extending into the prevertebral soft tissues bilaterally (labeled L and R) and appearing to extend posteriorly and to be in contact with the spinal canal (labeled ∗). Figure 3.15b is a double-contrast CT myelogram which shows the central portion of the MRI-enhanced mass (∗) to be a pseudomeningiocele which is sandwiched between the low attenuation left (L) and right (R) lateral masses. The CT finding dramatically changed the target volume and, consequently, the plan of treatment.

POSITRON-EMISSION TOMOGRAPHY (PET)

Positron emission tomography (PET) measures the three-dimensional distribution of a positron-emitting radioactive isotope within the body. Positrons are the positively charged antiparticles of electrons. They have the special property that, when they encounter an electron the two particles annihilate one another and emit a pair of gamma rays of equal energy (0.511 MeV each) and moving in opposite directions, back-to-back. Thus, a positron emitted by a radioactive isotope annihilates with an electron in close proximity to its point of origin and, if the locations of both of the emitted gamma rays are detected in an array of photosensitive detectors surrounding the patient, one can deduce that the original positron-emitting nucleus must have lain closely on the line between the two detected locations. By detecting a large number of such lines, the spatial distribution of the radioactive isotope can be deduced using a mathematical algorithm which is conceptually the same as that used in CT reconstruction.[6]

The main isotopes used in PET are: ^{11}C, ^{15}O, and ^{18}F, and there are several others. They are all relatively short lived, which poses a technical but not a fundamental problem. The unique and valuable aspect of PET is that these isotopes can be incorporated into molecules with specific biological properties which can be injected to concentrate in various body compartments. Typically, one can design tracers to identify concentrations of cells: with a high rate of metabolism; which are rapidly proliferating; or which are hypoxic.

---

[6] The reconstruction is complicated by the fact that some of the emitted gamma rays are lost by attenuation, so corrections have to be made to account for this.

Figure 3.16 shows, courtesy of V. Gregoire, a triplet of CT, MRI (T2-weighted) and ^{18}FDG axial scans at the same level of a patient with a hypopharyngeal tumor. An otherwise poorly appreciated right-sided node (labeled N) was seen in the PET study. In addition, the tumor size seen in the PET image (labeled T), and assessed by automatic contouring, was used to define the target volume which was smaller than would have been drawn using either the CT or MRI studies.

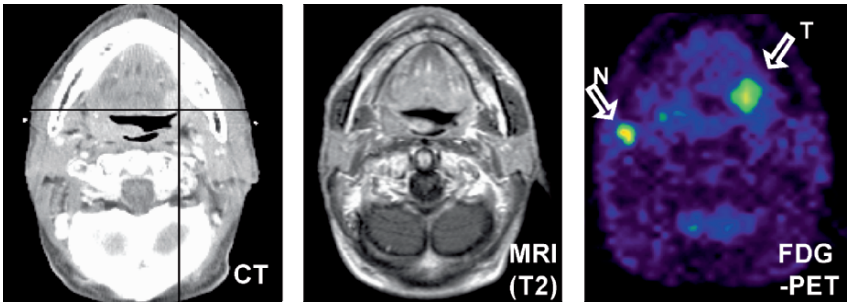


Figure 3.16. CT, MRI(T2) and FDG-PET studies (see text). Figure courtesy of V. Gregoire, UCL, Belgium.

PET images are spatially accurate since they share with CT the fact that position accuracy depends on purely mechanical features. However, their anatomic detail is poor and one needs, in addition, an anatomic map on which the PET-deduced activity can be super-imposed to ascertain precisely where a given locus of activity lies. While such a map can be obtained by inter-registration of the PET study with, say, a separate CT study, a combined-function CT/PET imager has been developed which makes the inter-registration issue trivial.

PET has two problems, both related to spatial issues. First, the spatial resolution of PET compared with CT and MRI is relatively poor; typically several millimeters. More problematic is that it is very hard to associate the activity measured in a PET image with a precise volume. By manipulating the way an image is viewed (that is, the mapping of activity to color), the apparent size of a hypoactive region can be altered significantly – making it hard to use PET to accurately establish, say, the extent of a target volume. Objective automatic methods of delineating regions of high or low activity have been developed. They are no doubt much more reproducible than manual contouring, but one is left with misgivings about the use of PET images to determine tumor boundaries accurately. Of course, all

imaging modalities are subject to the same sort of problem, but to a much lesser extent in the case of CT and MRI.

IMAGE REGISTRATION

Both Figure 3.14 and Figure 3.15, above, show side-by-side CT and MR images and, for these cases, such a presentation is adequate to spatially correlate the two studies and to extract the clinical information of interest. However, for more subtle comparisons one wants to be able to accurately correlate points in one image with points in the other so as to be able to combine the information from two studies in a spatially accurate manner.[7] This process is referred to as image registration or, alternatively, as image fusion. This topic has been been reviewed by, *inter alia*, Maintz and Viergever (1998) and by Kessler (2006). I first discuss image registration assuming the two images data sets are spatially accurate representations of a rigid body, and then for the situation in which one or both of the images are spatially distorted, or in which the patient has changed shape between the two studies.

Rigid body image registration

Consider first the inter-registration of two 2D sections such as a pair of transverse slices, either a CT and MRI section, or two CT sections, or a radiograph and a DRR. In the case of two 2D images, one might think that one could simply superimpose two semitransparent representations of the images upon one another and slide one over the other until they "matched." This was, indeed, the first thing I tried, many years ago. However, this does not in general

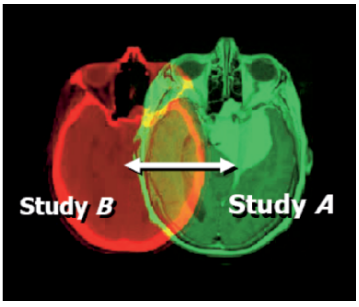


Figure 3.17. The problem of matching two disparate studies. Figure courtesy of M. Kessler, Madison, USA.

[7] Mathematically, one needs to compute the transformation matrix of translations and rotations (4 degrees of freedom in 2D, 6 in 3D) and, potentially, scale factors in each dimension. Of course, there may be more than two studies of interest. In that case, one can inter-register two studies, then additional studies of interest can be pair-wise inter-registered and hence any two studies can be related to one another.

work well – even if, as in Figure 3.17, one colors the two images differently (e.g., red and green, so that the overlap regions appear yellow). In the special case of two images of the same modality, the same orientation and the same scale, this registration technique is not too difficult. But, as is almost always the case, if the images have different scales, or are rotated relative to one another, or are from different modalities (e.g., CT and MRI, or radiograph and DRR), this approach is near-hopeless. The overlapped image is simply too confusing to the eye – and there are too many interacting variables to make accurate manual registration feasible.

In practice, there are three main approaches to rigid body image registration in both 2D and 3D, as now described.

Point-to-point registration This approach is based on the ability to identify of a pair of points, one in each section, which mark the locations of the same anatomic or fiducial feature. This is done multiple times for multiple anatomic features. If at least three non-collinear point pairs are identified, then the rigid body transformation between the two studies, namely the translations and rotations needed to bring them into alignment, can be calculated mathematically. That is the technical way of saying that, given the location of a feature in one image, one can compute the location of the same feature in the other image. While three non-collinear point-pairs are mathematically sufficient for reconstruction, the solution is much more robust – that is, is much less sensitive to errors in feature localization – if a larger number of features are identified and a least-squares fit is made to all the point-pairs.

Surface-to-surface registration There is a very profound problem with point-to-point registration. Namely, the human body really doesn't have anatomically distinct "points." It is composed of volumes of soft tissue or bone whose boundaries are delimited by surfaces, not points. The idea that anatomic *points* can be identified is intrinsically wrong and the attempt to do so is prone to error. A much sounder approach is to match anatomic *surfaces* with one another. One of the first surface-matching algorithms, developed in the early days of automated image registration, was the so-called "hat and head" model which matched the inner table of the skull in a pair of imaging studies (Pelizzari *et al.*, 1989). In the case of inter-registering a pair of 2D projection images, the surfaces of the various volumes such as skin, bone, airways, and so forth appear as *curves* and the task is to match pairs of curves with one another. This may

be done manually since the appearance of overlapped curves, or of a curve on top of an image, is not visually confusing. Matching can, of course, also be done automatically. Curve-to-curve matching has been used extensively in the 2D alignment of radiographs with DRRs.

Voxel-to-voxel matching Surface-to-surface matching requires delineation of the surfaces of the volumes of interest. This can be a demanding and labor-intensive step. Moreover, only a limited part of the information in the images is used in the matching process. A third approach is to match the image values at each point – that is, at each voxel in 3D or at each pixel (picture element) in 2D. For images of the same modality, an autocorrelation approach can be taken. More generally, the method of maximization of "mutual information" (Viola and Wells, 1995) has met with great success. One problem with voxel-to-voxel registration is that one tends to look at all the information in the images whereas, sometimes, certain parts of the image may be unreliable. For example, the mandible's location relative to the skull may be different at the times that two images were made, but it is irrelevant if one is only interested in the inter-registration of features within the skull; in that case, the portions of the images in which the mandible appears should be "thrown away." This is simple in theory, but very time-consuming in practice – whereas, with manual identification of points and/or surfaces, the selection of relevant anatomy is easy and instinctive.

When matching 3D data sets, one has to remember that there may, and generally will, be rotations and translations in the third dimension such that one cannot pair-wise match 2D sections. One should use a fully 3D approach.

Deformable image registration

The preceding methods can readily handle changes in scale between the registered studies by simply including scaling factors as variables to be determined in the fitting procedure. However, a much more vexing and difficult problem is when one study is spatially deformed relative to the other. Deformation may occur because the studies were taken at different times (e.g., on different days, or at different points in the respiratory cycle) or because one study is intrinsically spatially distorted (as may be the case with MRI). At the time of writing, the registration of deformed image sets is a matter of current research and there is no well-accepted solution. Figure 3.18 shows a comparison of rigid and deformable image registration. Figure 3.18a shows a CT cross-section of a prostate cancer patient's anatomy

during treatment simulation. The contours of the proximal seminal vesicle (part of the target volume) is shown in orange; the bladder is shown in blue; the rectum is shown in green; and the femoral heads are shown in purple. Figure 3.18b shows a CT image of the same patient during one of the treatment sessions. The CT image was acquired using an in-treatment-room CT-on-rails. It can be seen that the rectal gas and bladder filling have changed the anatomy. The contours overlaid on the CT image after a rigid-body image registration of the pelvic bones could not match the patient's anatomy. Figure 3.18c shows the same set of contours after performing a deformable image registration. It is clear to the eye that these contours match the anatomy much better than in Figure 3.18b. This particular automatic deformable image registration technique was developed by Wang *et al.* (2005).

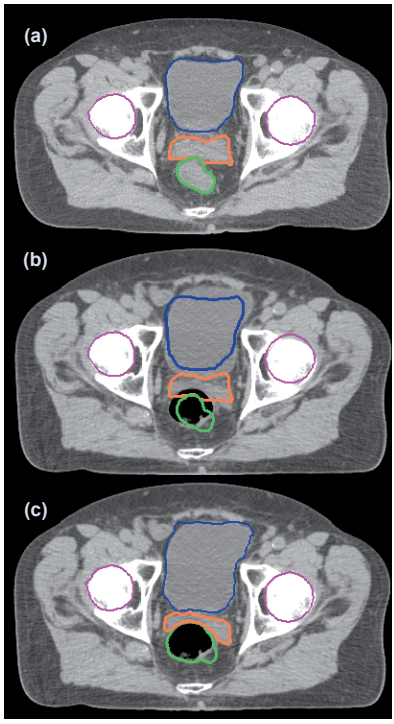


Figure 3.18. Comparison of rigid and deformable registration of contours (see text). Figure courtesy of Lei Dong, MD Anderson Cancer Center, USA.

The use of diagnostic imaging studies

Diagnostic studies are usually done: with the patient not in the treatment position; with only a partial field of view; and with possibly spatially distorted images. The emphasis on treatment accuracy has given rise to a mistaken impression that diagnostic studies may not be useful in designing the geometric aspects of a treatment. However, geometrically inaccurate studies often contain a wealth of valuable information which can be transferred to the planning CT. This might be done by deformable registration. But, it may often be done by visual inspection, using simple common sense. If, say, a diagnostic study shows a tumor fixed to C3 and extending down to mid-C4, up to the top of C2, and anteriorly by 3 cm, then this information can be manually transferred to the planning CT accurately and with ease.

DELINEATION OF ANATOMY

In order to be able to plan a radiation treatment, the relevant volumes of interest need to be "visible" to the planner. For this, it is generally necessary that they be identified in some way; that is, delineated.[8] It is possible, in principle, to treat a patient without ever explicitly identifying any volume of interest at all; beam shapes may be designed directly upon radiographs taken from the planned beam directions. However, to take full advantage of imaging information, it is normal to identify the target volume(s) and uninvolved normal tissues by delineating them in 3D with respect to an imaging study (or studies, in the case that multiple imaging modalities are used). There are two ways to delineate volumes of interest:

Manual delineation

The most common delineation approach is to draw contours following the outline of the feature of interest on sequential sections of an imaging study being viewed on a computer screen. Manual delineation on a computer is time-consuming when many sections and multiple features are involved.[9,10] However, when many finely spaced sections are available, it may not be necessary to delineate a feature manually on every section; interpolation between sections is possible and useful (Goitein and Abrams, 1983). The other problem with manual delineation is that it is subjective and error-prone; different observers may draw the volume of interest differently, and a given observer may draw the volume of interest differently on different occasions. An example of both of intra- and inter-observer non-reproducibility is given in Figure 3.19.

---

[8] The term "delineate" means to "describe or portray something precisely" (OED, 2001) and is not restricted to the manual drawing of outlines over an image.

[9] The term "feature" is widely used in computer-based delineation to identify an object of interest (e.g., a tank in military applications, organs and tissues in radiation therapy).

[10] It is a constant source of amazement to me that it is so hard to draw on a computer using a mouse or trackball. Writing one's signature accurately, for example, is almost impossible. A pen-like probe passing over a touch screen seems the best current solution, but it is still prone to parallax errors which are problematic for accurate work.

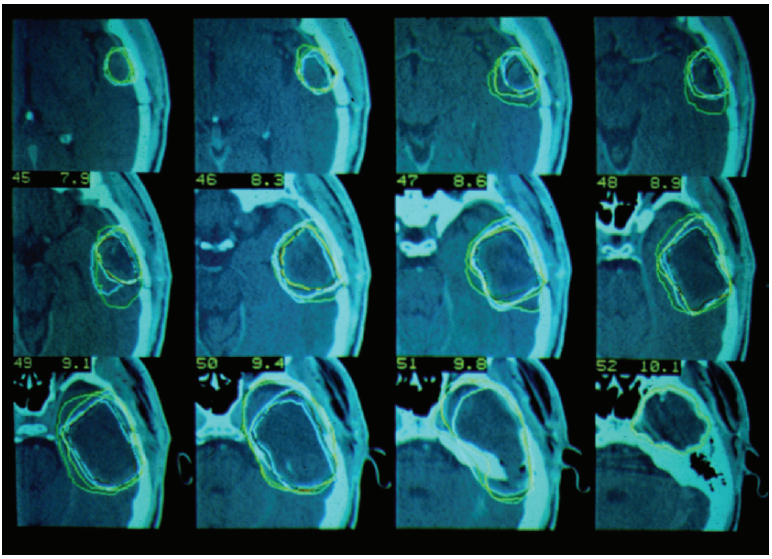


Figure 3.19. Delineation of the CTV on several sections of one patient with two delineations at separate times by two radiation oncologists (4 contours per section). Unpublished study by D. Pontvert and N. Liebsch, MGH, USA.

In this exercise, two physicians drew a target volume on each of two occasions on each of eight patients (only one of whom is represented in Figure 3.19).

Automatic feature extraction

There is much research into what is termed "automatic feature extraction." Currently, good success is achieved for high-contrast objects such as the external skin surface and outlines of the lungs, both of which involve high-contrast tissue/air interfaces, and for bone which is demarcated by the high contrast bone/soft tissue interface. However, the majority of features have much lesser contrast relative to their surroundings and, to date, there has been limited success in the reliable automatic extraction of most features of interest.

I mentioned in the introduction to this chapter that the delineation of the *tumor* is almost never possible by automatic means. The reasons for this are stated in the introduction and there is no need to repeat them here.

Uncertainty in the delineation of the target volume

The delineation of the tumor is subject to uncertainty and can be quite non-reproducible as was demonstrated in Figure 3.19 above. This non-reproducibility troubles many people – and one cannot say that it is desirable. However, in this connection, I want to make two points:

First, much of the variability seen in Figure 3.19 is, I believe, less a matter of inconsistency than a consequence of genuine uncertainties as to precisely where gross and subclinical disease is present. The fault, one might say, is in asking clinicians to draw a single sharp line. We would do much better to express our uncertainties explicitly.

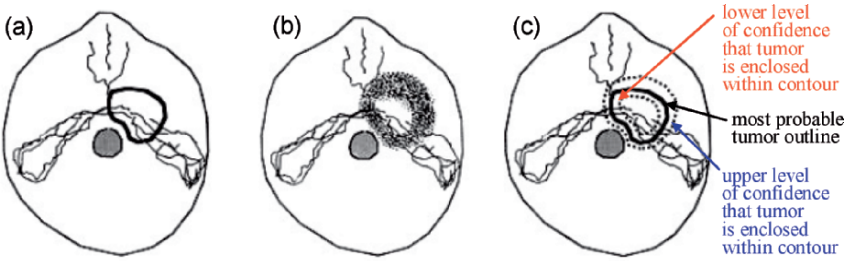


Figure 3.20. Three approaches to delineating a target volume: (a) traditional; (b) as a fuzzy boundary; and (c) as a most likely surface, with a confidence band around it (in 3D).

Figure 3.20 suggests a couple of ways to make the uncertainties in delineation both explicit and quantitative, as was also proposed by by Waschek *et al.* (1997). Panel (a) is the traditional single line delineation. In panel (b), a spray-can like tool has been used to express the delineator's guess at the probability that the surface of the volume passes through at a certain point (the probability density function). In panel (c), a triplet of contours has been drawn defining: the most likely outline; the outline outside of which the therapist is confident at a defined confidence level that there is no disease; and an outline within which he or she is confident, at the same level of confidence, that there is disease. Often the margin will not be the same all around the tumor, and may well be different for the upper and lower confidence contours. If we were to delineate out target volumes in such a manner, I would be willing to wager that there would be much more overlap between observers and for repeat delineations by a single observer than is suggested by Figure 3.19.

The second point relates to the fact that many people respond to the demonstrated uncertainties in target volume delineation by

concluding that there is no point in delineating volumes of interest to the last millimeter or so. I believe this is a wrong understanding of the matter. Very often it is not the appreciation of gross tumor *per se* which defines the surface of the GTV and/or CTV. Rather, it is the judgment that some normal tissue boundary sets a limit to tumor extension, or defines a volume which is to be excluded from the high dose region (e.g., the optic chiasm in the treatment of a base-of-skull sarcoma). This is the case even in Figure 3.19 where the observers have clearly made the judgment that the CTV does not extend into, or beyond, the inner table of the skull and in that neighborhood all four contours are highly congruent. Such constraints on the target volume due to adjacent normal tissues can often be delineated very accurately and are often the basis for spatially accurate and tight target volume delineation.

Delineation of uninvolved normal tissues and organs

The body contains hundreds if not thousands of anatomic structures. It is, of course, neither practical nor necessary to delineate them all. Usually the radiation oncologist will identify those structures which may be important in designing and characterizing the treatment. Even then, it is not at all uncommon to require in excess of a dozen structures to be delineated. This means that manual drawing can be very burdensome and this is the main impetus for the current interest in finding ways to extract anatomic structures automatically.

One attractive approach to what is called automated "feature extraction" is the use of a digital atlas of normal anatomy. In this method, a prototypical patient is created with a large number of anatomic features already delineated. This atlas must then be "fit" to the anatomy of the given patient. There is no problem, in principle, in creating such an atlas. It is time-consuming, of course, but need only be done once. (Actually, more than once, since one needs different atlases for males and females and, probably for very differently sized or developed individuals – fat and thin, child and adult.) There are two classes of problems encountered in the process of matching an atlas to a patient. First, a deformable registration will certainly be required, with the associated problem of how to match the atlas information to the information contained in the images. Second, the normal anatomy is very likely to be distorted by the tumor. The tumor may replace parts of the normal anatomy, and may displace and distort normal anatomy. These effects will, of course, not be

reproduced in the atlas. The problem is a difficult one, and a good solution is eagerly awaited.

SUMMARY

The ability to map anatomy is vital to planning radiation therapy. For decades, the only tools for this human cartography were the therapist's ability to: look with his or her eyes; feel with his or her fingers; and inspect X-ray radiographs which showed a tangle of superimposed anatomy with poor contrast. We are amazingly fortunate that a number of new imaging tools have dramatically changed the situation. CT, MRI, PET, and ultrasound have dramatically improved our ability to delineate and locate in space both gross tumor and the patient's normal anatomy. This ability is the *sine qua non* of modern radiation therapy.

4. DESIGNING A TREATMENT BEAM

INTRODUCTION

Just as a craftsman must know his or her tools, a radiation oncologist or medical physicist must know his or her radiations. So, in anticipation of the discussion in Chapters 8 and 9 of how to design a radiation treatment using multiple photon beams, I want to present some simple aspects of the interactions of radiation with matter which explain the properties of a single beam of radiation. These interactions are not treated in detail; much more comprehensive accounts can be found in textbooks such as Johns and Cunningham (1983) and (Khan 2003).

A number of radiations have been used in the treatment of cancer, among them: X- and γ-rays, electrons, neutrons, protons, α-particles, and heavier ions such as carbon and neon. Of these, by far the more commonly used nowadays are photons produced by an electron linear accelerator. The physics underlying photon therapy will be discussed in this chapter, and that underlying proton beam therapy in Chapter 10. While they all have great interest, I will not discuss the therapeutic use of the other radiations. However, as we will soon see, one cannot talk about the physics underlying either photon or proton beam therapy without considering the interactions of electrons with matter.

The discussion of the interactions of photons with matter is presented in two parts: first, the manner in which an individual photon interacts with an individual atom; and, second, the manner in which a photon beam (comprised of many photons) interacts with matter (with its many atoms and molecules).

THE INTERACTIONS OF PHOTONS WITH INDIVIDUAL ATOMS

Photons are rather mysterious. They may be thought of as electromagnetic waves propagating through space (originally described as a disturbance of the ether) – electromagnetic in the sense that they are accompanied by oscillating electrical and magnetic fields. Or, they may be thought of as quasi-particles – chargeless packets of energy which travel at the speed of light, equal to $3.0 \cdot 10^8$ m·s^{-1}, through space. This wave-particle duality is one of the mysteries of physics – understood by very clever physicists and a bit of a puzzle to the rest of us. For the present purposes, I will stay with the particle description and use the term *photon* for both X-rays and γ-rays to characterize one such packet of energy. Energy is typically expressed in units of *electron volts* where one electron volt (abbreviated eV) is the amount of kinetic energy gained by a single unbound electron when it passes through an electrostatic potential difference of one volt, in vacuum. Visible light has an energy ranging from a bit less than 2 eV (red light) to about 3 eV (blue light), photons used for diagnostic purposes (e.g. for radiographs) have a range of energies from about 50 to 150 keV (thousands of eV), and modern therapeutic photons have a range of energies from about 1 to 20 MeV (millions of eV).

When a photon passes in the vicinity of an atom, its electromagnetic fields exert forces on the positively charged nucleus and on the

negatively charged orbiting electrons and, in the extreme case, these forces are strong enough to tear the atom apart and, hence, cause biological damage. This interaction can occur in one of three main ways which I now briefly discuss. For each of these, I will address two issues, namely the energetics of the interaction and the angular distribution(s) of the resulting product(s). These are both constrained by two extremely powerful laws of physics, namely the law of conservation of energy, and the law of conservation of momentum.

Photo-electric interactions

Einstein is probably best known for his theory of special relativity, but his 1921 Nobel prize was won for his explanation of the photo-electric effect, given in a paper entitled *On a Heuristic Viewpoint Concerning the Production and Transformation of Light*. In this paper he developed the notion of the photon and showed how a photon could "collide" with an electron, transferring all of its energy to it. In an atom, some of this energy is needed to overcome the so-called binding energy of the electron within the atom, the remainder is transferred as kinetic energy to the electron that then escapes, leaving behind an ionized atom, as sketched in Figure 4.1. The ionized atom is likely to undergo further changes as its electrons readjust themselves in their orbits. This can result in the emission of relatively low energy photons or electrons – but I will not discuss these secondary consequences further.

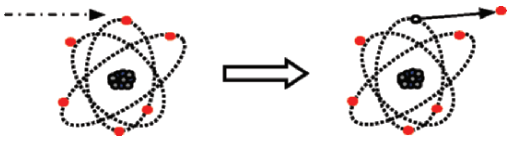


Figure 4.1. Schematic representation of a photo-electric interaction. In this and subsequent similar figures, electrons are represented as red dots, the photon is represented by a dot-dash line, and the path of the ejected electron by a continuous line. An unfilled dot represents the location from which an electron has been ejected.

The law of conservation of energy leads to the simple expression:

$$E'_{electron} = E_{photon} - E_b$$

where E_{photon} is the energy of the incident photon, $E'_{electron}$ is the energy of the ejected electron, and E_b is the energy with which the electron was initially bound to the atom.

So far as the angular distribution of the ejected electrons is concerned, at the energies of therapeutic interest, the electron tends to be ejected in the near-forward direction,[1] although they do have a low probability of being ejected in the backward direction.

Compton interactions

In contradistinction to the photo-electric effect, where the photon energy is given up in its entirety and the photon disappears, in the Compton effect, the impinging photon imparts only some of its energy to an orbiting electron within an atom, and continues on with reduced energy (in, typically, the near-forward direction). The photon is said to be scattered. The target electron usually receives sufficient energy to be ejected from the atom, typically at a finite but small and opposite angle from the scattered photon's direction as sketched in Figure 4.2.

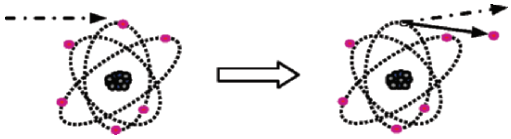


Figure 4.2. Schematic representation of a Compton interaction of a photon with an atomic electron.

The law of conservation of energy leads to the relationship

$$E'_{electron} + E'_{photon} = E_{photon} - E_b$$

where E_{photon} and E'_{photon} are, respectively, the energies of the initial and scattered photons, $E'_{electron}$ is the energy of the ejected electron, and E_b is the energy with which the electron was initially bound to the atom. The scattered photon and ejected electron share the available energy with one another. The photon, for example, may carry off anywhere between almost all to almost none of the available energy, and the electron would correspondingly carry off anywhere between almost none to almost all of the available energy.

Photons can be scattered over the full 4π range of angles. The more energetic the incident photons, the more forward-peaked is the angular distribution of scattered photons. Over half of the scattered photons from a 4 MeV incident photon will be within about ± 10° of

[1] By the "near-forward direction" I mean an angle that is within about 10 to 15 degrees of the direction of the initial photon.

the forward direction. The electrons are always ejected within $\pm 90°$ of the near-forward direction.

A Compton interaction leaves behind an ionized atom which, as with photo-electric interactions, is likely to undergo further changes as its electrons readjust themselves in their orbits – such as emitting relatively low energy photons or electrons.

Pair production

The third principal interaction of photons is in near-collisions with the atomic *nucleus*, which lead to extinction of the photon and the creation of a pair of particles: an electron and a positron (which is the anti-particle of the electron) as sketched in Figure 4.3. This Process, termed pair production, is a dramatic example of Einstein's theory of the equivalence of energy and mass. Here, the energy of an incident photon is transformed to create two particles, each of mass 0.511 MeV. As a consequence, there is a threshold energy for pair production, namely $2 \cdot 0.511$ MeV = 1.02 MeV. A photon of lesser energy simply does not have the energy needed to produce a particle/anti-particle pair. The energy relationship is

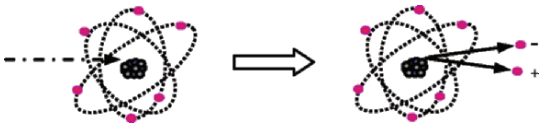


Figure 4.3. Schematic representation of pair production. An electron and a positron (positive charge) are created in the field of the atomic nucleus.

$$E'_{electron} + E'_{positron} = E_{photon} - 1.02\ MeV$$

where E_{photon} is the energy of the incident photon, and $E'_{electron}$ and $E'_{positron}$, and E_{photon} are the energies, respectively, of the emitted electron and positron. The electron and positron share the available energy. They are produced at an angle of typically 0.511/T radians, where T is the kinetic energy of the particle in MeV. Thus, very approximately, a 4 MeV photon will produce electrons and positrons half of which will lie within about $\pm 15°$ of the forward direction.

Dependence of photon interactions on photon energy

Figure 4.4 shows how the relative importance of each of the three main modes of photon interaction varies with the energy of the photon. The likelihood of a photo-electric interaction varies roughly as $(1/E_{photon})^3$. As a result, photo-electric interactions are much more important – indeed, are dominant – for low energy photons. Above

about 0.05 MeV, the relative importance of Compton scattering is roughly constant with energy up to 3 MeV or so and then falls slowly. The likelihood of pair production is zero up to a threshold of two electron masses, and then rises sharply with energy. As a result, pair production is much more important – indeed, is dominant – for high energy photons. However, the range of therapeutic beam energies is from about 0.3 to 20 MeV and, in that range, as Figure 4.4 demonstrates, the Compton interactions totally dominate the picture. *Thus, the behavior of most therapeutic photon beams is explained by the physics of Compton interactions alone.*

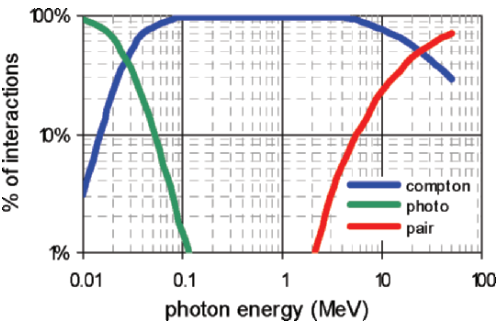


Figure 4.4. Graph showing the relative importance of photoelectric effect, Compton interactions, and pair production in water. *N.B.* both axes are logarithmic.

Dependence of photon interactions on atomic number (Z)

The likelihoods of the three main interactions of photons have very different dependencies on the atomic number (Z) of the target atom. Per gm·cm^{-2} of material, the probability of an interaction is: for photoelectric interaction, approximately proportional to Z^3; for Compton interactions, virtually independent of Z; and, for pair-production, approximately proportional to Z. Thus, for example, at diagnostic energies (about 0.1 MeV or less) the photoelectric effect in elevated-Z materials such as bone is very important – which is why such good bone contrast is possible in diagnostic radiographs. The domains of dominance of the three interactions as a function of both photon energy and target material are shown in Figure 4.5. This figure confirms that, except perhaps

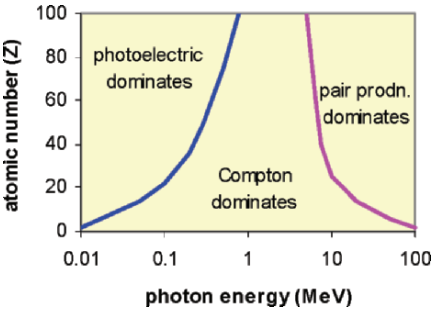


Figure 4.5. Domains of dominance as a function of photon energy and atomic number

for metallic objects within the patient, the behavior of therapeutic photon beams in all body parts is dominated by the physics of Compton interactions.

Interactions with molecules

So far, the discussion has been confined to interactions of radiation with atoms. Human tissue is, of course, made up primarily of molecules – which themselves are constituted of atoms. By and large, it is the interactions of radiation with atoms that are fundamental. However, once an atom has been ionized or excited, its parent molecule will be affected. If the atom has been ionized, some of the bonds linking atoms within the molecule are likely to be disrupted and the molecule thereby either broken up or substantially transformed. Water, for example, which constitutes some 60% of our body, may be transformed into free radicals which themselves are highly chemo-active. The molecule may also take up energy through rotational and vibrational excitations which ultimately appear as heat.

As we will soon see, the electrons set loose by the interactions of photons with individual atoms play an important part in the way in which photons cause dose to be deposited in bulk matter. So, we must take a moment to address the electron's interactions.

THE INTERACTIONS OF ELECTRONS WITH INDIVIDUAL ATOMS

Just as photons exert forces on the constituents of atoms, so do electrons. However, the mechanism is rather different, since electrons carry a charge ($1.60 \cdot 10^{-19}$ Coulomb per electron) while photons are neutral. Any two charged objects, even when not moving, exert an equal and opposite force on one another – called the Coulomb force.[2] The force is attractive when the objects are oppositely charged and repulsive when they carry the same charge. The force is inversely proportional to the square of the distance between the objects.

Thus, when an electron passes near to or through an atom, it exerts a force on the orbiting electrons and on the nucleus, giving rise to four

---

[2] Charles-Augustin de Coulomb was a French physicist working in the second half of the 18th century who made many discoveries in the fields of mechanics and of electricity and magnetism.

types of interaction: excitation, ionization, scattering, and bremsstrahlung. The first two arise from interactions of the incident electron with orbiting electrons; the third and fourth from its interactions with the nucleus of the atom.

Excitation

The incoming electron may transfer energy to one or more of the atom's orbiting electrons, which are then "excited" – for example, forced into another orbit. As these electrons then redistribute, they can both impart energy to the atom as a whole which shows up as heat, or they can induce secondary radiations, both soft photons and low energy electrons. These secondary radiations are not responsible for much damage to tissues. However, the incident electron will continue on, with its energy only slightly diminished (by an amount equal to the excitation energy of the atom) and will experience further interactions which themselves are highly likely to cause damage.

Ionization

When the energy transferred by the incoming electron exceeds the binding energy of a target electron, that electron will be ejected from the atom, which will be left in an ionized state, potentially causing biological damage. The incident electron will lose energy, and continue on. Figure 4.6 is a schematic representation of a Coulomb interaction of an electron – and bears an obvious similarity to Figure 4.2 which illustrates a Coulomb interaction of a photon. The principle of conservation of energy implies that

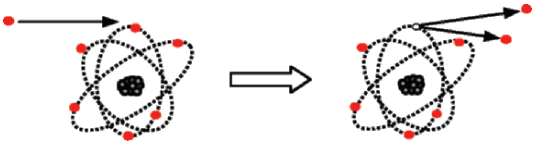


Figure 4.6. Schematic representation of an ionization resulting from the Coulomb force between the incoming electron and orbiting atomic electrons.

$$E'_{electron} + E''_{electron} = E_{electron} - E_b$$

where $E'_{electron}$ and $E''_{electron}$ are the energies of the two final electrons, $E_{electron}$ is the energy of the incident electron, and E_b is the binding energy of the ejected electron. There is a subtlety here. One cannot distinguish between the two ejected electrons in the sense of knowing which is the original incident electron and which is the ejected electron.

There is a traditional, dating back to the days of the use of cloud chambers, and somewhat vague, nomenclature used to describe the electrons resulting from ionization. If an electron resulting from an interaction has sufficient energy to cause further ionizations, it is termed a *delta-ray*, usually written as *δ-ray*. Typically, δ-rays have energies above some 10 to 30 eV.

The ionized atom, as always, is likely to undergo further changes as its electrons readjust themselves in their orbits – such as emitting relatively low energy photons or electrons.

Scattering by nuclei

Electrons can also be scattered by the nuclei of atoms as shown schematically in Figure 4.7. So far as deflection of the electron is concerned, this affects the electron's deflection much more than scattering by atomic electrons. Although each interaction with a nucleus will cause only a slight deflection, the electron will have a very large number of such interactions before it comes to rest, and the accumulation of deflections can cause the electron path to deviate greatly from a straight line. In fact, an electron can be so greatly deflected by the accumulation of scattering events that it can be "turned around" and end up traveling in the backwards direction.

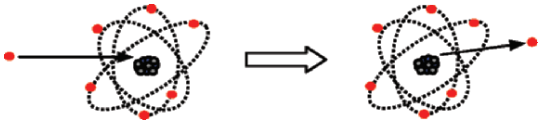


Figure 4.7. Schematic representation of an electron scattering off the nucleus of an atom.

Bremsstrahlung

When electrons are accelerated, they give off electromagnetic radiation. For example, the electromagnetic waves used to transmit radio broadcasts are created by causing electrons to oscillate, and therefore be accelerated, within an antenna. When an electron passes near an atomic nucleus it experiences a powerful sideways acceleration. This causes the emission of electromagnetic

Figure 4.8. Schematic representation of bremsstrahlung. A photon is created as the electron is deflected by the charge of the atomic nucleus.

radiation (i.e., of photons) – and, also, deflection of the electron. Figure 4.8 schematically depicts this process. The law of conservation of energy implies (ignoring the small amount of energy taken up by the recoiling nucleus) that

$$E'_{electron} + E'_{photon} = E_{electron}$$

where $E_{electron}$ is the energy of the incident electron, $E'_{electron}$ is the energy of the ongoing electron, and E'_{photon} is the energy of the emitted photon. Both the electron and the photon tend to go off in the near-forward direction; the higher the energy, the more forward the trajectory. The energy released in the bremsstrahlung process is proportional to the first power of the atomic number, Z, of the atom.

There is a curious aspect of the bremmstrahlung process, namely that detailed calculations of the effect predict the emission of an infinite number of photons! Physicists hate infinite values, and this was quite a puzzle when it was first realized. However, the solution was fairly soon forthcoming. It transpires that, while there are indeed an infinite number of photons emitted, most tend to carry off an infinitesimal amount of energy. As a result, the net energy carried off by all photons in any fixed range of energies, down even to zero, is finite – so, in this manner, the infinite problem is resolved.

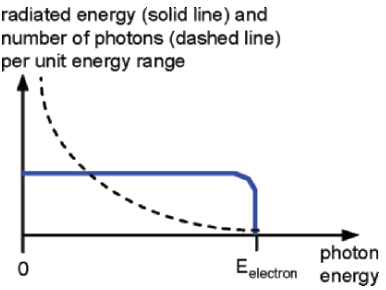


Figure 4.9. Schematic representation of the photons resulting from bremsstrahlung of an electron beam of energy $E_{electron}$. At low photon energies, a near-infinite number of photons are produced (*dashed line*), but the energy emitted per unit energy is virtually constant with photon energy (*blue line*).

In addition, the net energy carried off by the photons in any fixed range of energies is more or less constant as a function of the photon energy. This is shown schematically in Figure 4.9.

While bremsstrahlung is an important effect in many applications (e.g., in the targets of X-ray tubes and linacs used in radiation therapy), it is not a substantial factor in the interactions of radiation with tissue, as we will discuss below.

THE INTERACTIONS OF PHOTONS WITH BULK MATTER

So far, we have mainly considered only single interactions with a single atom. Now it is time to look at what happens in matter, which is comprised of a whole lot of atoms. First, however, a few words about the concept of dose.

The concept of dose

When radiation interacts with matter, energy is lost and much of it is transferred to the atoms and molecules of the matter. The lost energy may be transferred at, or very close to, the site of an interaction, or it may be transferred some distance away from the site of the original interaction by secondary photons and particles. Some of the energy is carried away by photons and particles that exit the target material and can have no further impact upon it.

Dose is a measure of the amount of energy deposited in a small volume at a point of interest as a result of the radiation – be that energy deposited locally, or brought to the point of interest by secondary radiation generated at some distance from the primary interactions. Dose is expressed in units of Gray (written *Gy*), where one Gray is equal to 1 $J \cdot kg^{-1}$.[3]

In principle, one could measure the dose by measuring the energy deposited in a volume which is small with respect to the spatial variations in dose, and then dividing the measured energy by the measured mass of the material in the small volume. There are, as ever, all sorts of technical details connected with actually making such a measurement. Suffice it to say that the most common approach is to make measurements with a small ionization chamber[4]

---

[3] The Gray is an SI unit, primarily created for the purposes of radiotherapy. In times gone by, the unit of dose was the rad. Its definition was such that 1 rad is equal to one hundredth of a Gray – which may be written 1 cGy.

[4] An ionization chamber is a small cavity with generally a central electrode or a pair of parallel flat electrodes. Radiation passing through the cavity ionizes the gas, creating many ion–electron pairs, the number of which is proportional to the dose deposited by the radiation. A voltage is applied between the electrode and the cavity wall (which is made to be conducting) and this voltage gradient causes the electrons and ions to drift apart towards opposite electrodes. The charge carried by the electrons reaching the positive electrode is measured with an electrometer and the dose is equal to the charge collected multiplied by a calibration factor and by a

[continued on next page

placed at the point of interest, whose calibration, together with the electrometer with which it is read out, is traceable to a primary standards laboratory. In the United States the National Institute of Standards and Technology uses a water calorimeter – an instrument that measures the heat deposited in a unit mass of material – as its standard for absorbed dose in water from a ^{60}Co beam.

The question arises: In what medium should the measurement be made? Exposed to the same radiation, different materials receive slightly different doses at a given point. By convention, in radiotherapy, the dose is reported as though the medium were water.

What form does the deposited energy take? We have already sufficiently addressed the interactions of radiation with matter to be able to answer this question. In the end, *all the energy appears either as heat, or in the form of chemical changes resulting from ionizations*. In practice, by far the greatest proportion of the deposited energy – at least 96% – appears as heat. Anyone who has touched or leant against the head of a ^{60}Co therapy machine knows well the warmth generated in the shielding surrounding the source.[5] Of course, it is the chemical changes that lead to tissue damage, and I always find it a bit surprising that so small a fraction of the radiation's energy loss is therapeutically effective.

The experience of a single incident photon

Consider a 4 MeV photon[6] which impinges upon a patient – say, laterally in the brain where it potentially can pass through perhaps 14 cm of tissue and bone.

> *Question:* what is the most probable thing that will happen to that photon?
>
> *Answer:* absolutely nothing!

---

generally small temperature and pressure correction factor. (See, also, Chapter 10.)

[5] A patient exposed to therapeutic doses of radiation (e.g., 2 Gy) experiences a warming of his or her irradiated tissues. However, the temperature rise is of the order of some $5 \cdot 10^{-4}$ degrees and is imperceptible. (Actually, the temperature rise will be even less since this number is based on calculation in a static situation and, in practice, much of the heat is likely to be carried away by blood flow)

[6] The reason for picking 4 MeV will become clear shortly.

Yes, it is so. The interaction probability is so low that the most likely thing to happen is for the photon to pass through the patient and out the other side without suffering any interaction at all – and hence without doing any damage to the patient's tissues whatsoever.

The second most likely event is for the photon to have a single interaction. As we have seen, it is Compton scattering which dominates the interactions of photons in the therapeutic range, so that collision will just about certainly be a Compton interaction. That is, in the case, for example, of a 4 MeV photon, the scattered photon will continue on with diminished energy (anywhere from about 0 to 4 MeV), and an electron will be ejected from the target atom with an energy between ~4 and ~0 MeV. The most probable thing that will happen to the scattered photon is, again, nothing! It is likely to escape our patient with no further interactions.

We will come back to the scattered photon's fate in a moment, but let us now concentrate on the ejected electron – and let us imagine that it got about half of the available energy, i.e., ~2 MeV. What will it do? Well, unlike the neutral photon whose interaction probability is small, a charged particle like the electron has a very high probability of interacting. As we know, it will either excite or ionize atoms. In the first case, one, and in the second case, two electrons will emerge, still carrying a lot of energy – namely the full energy of the incident electron minus the binding energy of the ejected electron which is of the order of 10's to a few 100 eV. That is, very little energy will have been lost. So, the still energetic electron(s) continue on to have yet other interactions, and their children will have other interactions, and so on *ad* (nearly) *infinitum*. In the end, since binding energies tend to be of the order of tens of eV (say, 50 eV), there will typically have been about 2 MeV divided by 50 eV $\simeq$ 40,000 ionizations before the incident electron and its progeny lose all their energy and come to rest. This large number of interactions results in the electron loosing about 2 MeV per centimeter of water path creating a "splash" of dose in the neighborhood of the interaction of something like 1 cm length and a few millimeters width.

The upshot of all this is that, while the initial photon interaction will have ionized a single atom, the ejected electron will go on to ionize tens of thousands of atoms. That is, *virtually all the damage caused by a photon is due to damage caused by secondary electrons*.

Let us now return to the scattered photon that emerged from the initial Compton interaction. I had said that the most likely thing was that it

would escape the patient without further interactions. However, just as with the initial photon, there is a chance that it will itself suffer a further Compton interaction. If so, the interaction will probably be some distance away from the site of the initial interaction, since the interaction probability of photons is low. The second interaction will proceed very much like the first. There will be a scattered photon which will probably escape the patient without doing further damage and the ejected electron will cause 10's of thousands of ionizations as it loses its energy and comes to rest. The only difference is that the scattered photon will have less energy than the initial photon, and so the, on average, lower energy ejected electron will cause somewhat fewer ionizations over a slightly smaller distance than in the case of the first Compton interaction – that is, its "splash" of dose will be slightly smaller in extent. And, of course, the scattered photon from the second Compton interaction *may* suffer yet another Compton collision and so on… Figure 4.10 recaps this story. In this figure, scenario (a) is most likely; scenario (b) is the next most likely; and scenario (c) is the least likely to occur.

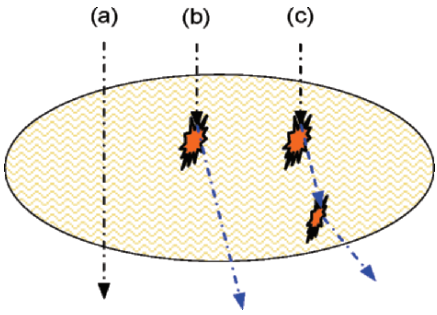


Figure 4.10. The biographies of 3 photons: (a) The photon passes through the patient without interaction; (b) the photon suffers one Coulomb interaction with the ejected electron causing 10's of thousands of ionizations (*red splash*); (c) the scattered photon from the first Compton interaction suffers a second Compton interaction, and its ejected electron causes 10's of thousands of further ionizations (second *red splash*);

The experience of many incident photons

Typically, for a beam of 4 MeV photons for example, several times 10^{10} photons per cm^2 are needed to deliver a dose of 2 Gy. A beam comprised of such a large number of photons will result in overlaying an equally large number of histories such as those depicted in Figure 4.10. As a result, there will be a large number of electron-induced dose "splashes" laid down all throughout the irradiated medium. These will overlap one another and will merge into one big dose "splash" extending throughout the volume of tissue which is within

the beam. In addition, there will be a "sea" of secondary photons that eventually escape the patient and, therefore, do him or her no further damage.

We now go on to more precisely analyze the nature of the dose distribution, and see how it can be "shaped" so as best to achieve one's objectives. But first, a very few comments on how therapeutic photon beams are generated.

THE GENERATION OF THERAPEUTIC PHOTON BEAMS

Radioactive isotopes are one source of radiation, and the ^{60}Co therapy machine takes advantage of this. A highly active source of ^{60}Co is placed in a heavy lead shield which has an aperture through which the photons produced in the decay of ^{60}Co can escape to provide the therapeutic beam. The whole is then usually mounted on a rotating gantry so that the beam can be directed at the patient from any angle. ^{60}Co therapy machines are little used these days, except in areas of the world where the supply of electricity and/or repair service are problematic. I mention these machines because they are unusual in that their photon beam is near mono-energetic. It consists primarily of γ-rays of 1.17 and 1.33 MeV energy – which are close enough together that one can think of the radiation as consisting of 1.25 MeV primary photons. However, photons interacting with the shielding around the ^{60}Co source produce lower energy secondary photons which lower the effective energy of the beam somewhat.

Most therapeutic beams today are produced by electron accelerators, abbreviated "linacs" (see Figure 1.1 in Chapter 1) which produce an intense beam of electrons. These electrons, when they impinge on a tungsten target[7], produce a beam of photons via the brems-strahlung process.[8] The main differences between such beams and

---

[7] You will recall that the probability of the bremsstrahlung process is proportional to Z. Tungsten is used because, as well its high density (and therefore compactness) and good heat toleration, it has a high atomic number.

[8] As a side note: for photon beams of therapeutic energies, the electron beam is pointed in the direction of the desired radiation beam since the bremsstrahlung process is forward-peaked at high energies. In diagnostic

[continued on next page

^{60}Co beams are threefold: their penumbra tends to be smaller since the effective source of a linac beam source is much smaller than the size of a ^{60}Co source; they can be more penetrating because the electron beam energy can be high (20 MeV or more); and, lastly, they are not at all monoenergetic. Rather, their energy spectrum tends to be somewhat similar to that shown in figure 4.9. Namely, dose is delivered over a large range of energies, extending up to the energy of the electron beam which created them but with the lowest energy photons filtered out. As a rule of thumb, the effective energy of a linac beam (the energy of a mono-energetic beam with similar depth-dose characteristics) is about 40% of the peak energy. This explains why I have been using 4 MeV photons to illustrate a number of things in this chapter. This is close to the effective energy of a 10 MeV linac photon beam – and 10 MeV linacs are widely used in current radiotherapy practice.

One final note. While the bremsstrahlung process produces a fairly broad beam, it nevertheless, tends to be forward-peaked so that, if no steps were taken, the flux of a linac's photons would be higher near the center of the field than at greater radii. To correct for this forward-peaking of the radiation, a conically-shaped *flattening filter*, thicker in the middle than at the edges, is interposed in the beam at some distance from the target in order to attenuate the forward-peaked photons preferentially. The shape of the flattening filter can be designed to produce a flat field at some chosen depth. However, this is at a cost. The filter tends to absorb more lower than higher energy photons; so, the beam is "harder" in the middle of a field, where the flattening filter is thickest, than at its edges. A consequence of this is that a beam can be made flat at only one depth. At shallower depths it will be "cupped" and, at larger depths, "humped." In the early days of linac therapy the spatial variation of the beam–hardening effect was not fully realized. Beams were flattened at depth (usually at 10 cm depth), but near the patient's skin surface elevated doses were delivered, especially at large field radii. The problem was compounded because people tended to make measurements of lateral profiles along the principal axes of square fields and omitted to look diagonally in the corners of the field which, as they extended to a larger radius, had even higher "horns". This led to some quite

---

X-ray tubes, the photon beam is selected at a sideways angle relative to the electron beam direction, since the bremsstrahlung process at low energies is peaked more nearly at 90° to the direction of the electrons.

undesirable patient reactions until the problem was appreciated and cured. This was one of the not-infrequent situations in which dosimetric problems were first discovered by clinicians, due to untoward patient reactions and underlines out how sensitive the patient is to the dose he or she receives.

THE DESIGN OF A UNIFORM RECTANGULAR TREATMENT BEAM

Let us jump ahead a little, and describe the formation of a simple beam of photons. Figure 4.11 depicts this schematically. In this figure, the beam delivery system (which is usually mounted on a rotating gantry) is comprised of everything within the trapezoid region, and the medium being irradiated is, in essence, a bucket of water.

For the moment, let us assume that the collimator is a square hole in a metal block, and that there is no patient-specific aperture or intensity-modifying device in the beam. I want to discuss now what the dose distribution looks like along the two dotted blue lines shown in the figure.

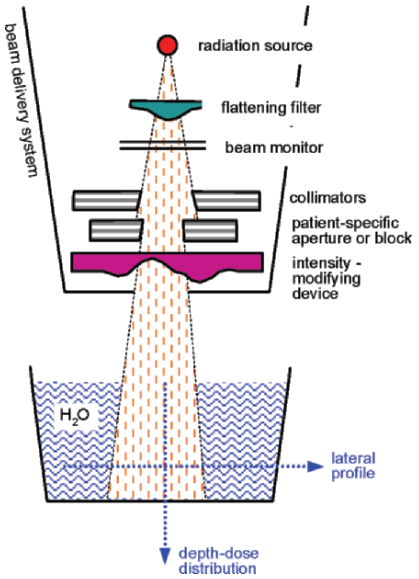


Figure 4.11. Schematic representation of a radiation beam impinging on a bucket of water.

Distribution of the dose in depth

Figure 4.12 shows a sequence of semi-logarithmic depth dose curves, taken along the central axis of the beam, in which various physical effects are "turned on" in successive panels of the figure. We now discuss these effects.

When photons impinge on matter, their number gets attenuated as depth increases, simply due to the loss of photons from upstream Compton interactions. And, as their number decreases, the dose that they deposit decreases proportionately. Less photons create less "splashes". This attenuation is, to a first approximation, exponential with depth. That is, one can write:

$$n = n_0 \, e^{-\mu . d}$$

where n_0 is the number of incident photons, n is the number of photons at depth d, and μ is a physical constant characteristic of the target material and a function of the photon energy, termed the linear attenuation coefficient. The property of exponential attenuation expresses a very important physical principle which applies in many other areas too, such as in radioactive decay. Exponential behavior occurs when:

- there is a set of a large number of objects (e.g., photons or radioactive atoms) each of which can experience some process (e.g., an interaction or a decay);
- once an object experiences the process in question, it is removed from the set of objects;
- the probability of a process occurring is independent of the occurrence of previous processes.

These conditions are only partially met in a photon beam, as we shall see. If we assume that the deposition of dose at a point is proportional to the number of photons at that point, and that exponential attenuation is taking place, then the dose distribution would be as schematically depicted in Figure 4.12a.

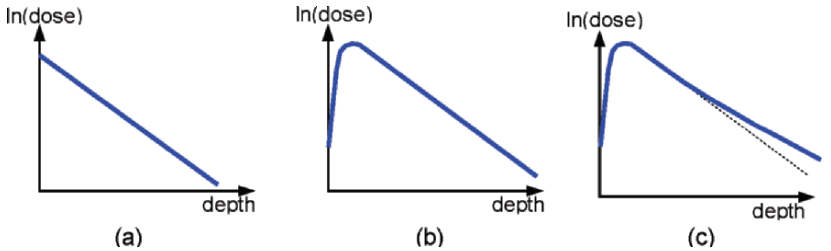


Figure 4.12. Schematic representation of the depth-dose distribution of a photon beam with dose plotted logarithmically. (a) First order approximation; (b) with build-up; and (c) with beam hardening and other effects (see text).

However, dose is *not* proportional to the number of photons at a point but, rather, to the energy deposited by the secondary electrons. You will remember that the secondary electrons travel of the order of 1 cm beyond the interaction point of, say, a 4 MeV photon – and, that they tend to travel in the forward direction. The number of secondary electrons builds-up below the surface of the irradiated medium, giving rise to dose distributions such as are illustrated in Figure 4.12b.

How this *dose build-up* comes about is schematically illustrated in Figure 4.13. In this figure, a set of photons (dot-dash arrows) impinge on a block of material and they have their first interactions at successively greater depths. Each then lets loose an electron (red arrow), which travels in the forward direction for a fixed distance and then stops. The number of electrons at a given depth in this example is shown for several depth levels. This number begins from zero, then increases incrementally until reaching an "equilibrium" value of six. In practice, of course, there is a huge number of photons, the electrons they produce have a range of energies, and the electrons do not all travel in the forward direction. While this complicates the picture slightly, the same principle applies: there is a gradual build-up of dose until an equilibrium level is reached at a depth strongly related to the average distance traveled by secondary electrons (e.g., a few millimeters up to a few centimeters for very high energy photons). The dose reaches a maximum value, due to the counter-acting effects of dose build-up and exponential photon attenuation.

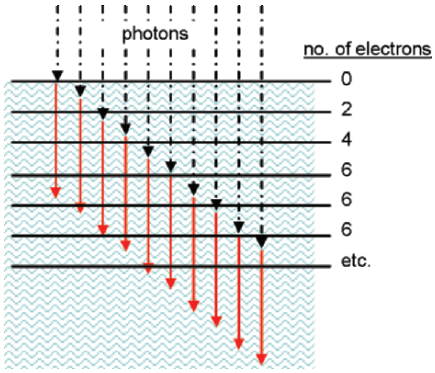


Figure 4.13. Schematic explanation of build-up effect (see text).

In this idealized example, the dose at the very surface of the material is essentially zero. The estimate of zero entrance dose is due to the assumption that all secondary electrons travel forward. In fact, the angular distribution of electrons from Coulomb interactions has a wide spread of angles, including a minority of electrons which, due to scattering effects, travel in the backward direction. These, together with contaminants in the beam such as electrons emanating from the collimator edges or material in the beam, raise the entrance dose to some tens of percent of the maximum dose. How much they raise it is a function of photon energy and field size.

As a result of the build-up phenomenon, high energy photons have a low entrance dose. In patients, this so-called *skin-sparing effect* is extremely important in avoiding high doses to the epidermis. In earlier times, when only low energy photon beams were available

(produced by so-called orthovoltage machines with a peak photon energy of less than 0.3 MeV and no skin-sparing property), skin reactions such as moist desquamation, telangiectasia, and skin necrosis strongly limited the dose which could be delivered to deep-seated tumors. The advent of ^{60}Co teletherapy machines in the late 1950's, with their skin-sparing properties and better depth penetration, revolutionized radiation therapy.

Three other effects cause deviations from exponential dose fall-off at larger depths, as depicted in Figure 4.12c, and now described.

Inverse-square fall-off When radiation emanates from a localized source, the intensity of radiation diminishes with the inverse square of the distance from the source. This is because the amount of radiation is constant, but the area of surface over which it is spread out increases with the square of the distance. Thus, even if there were no photon attenuation, the ratio of the dose at a point 10 cm below the surface to that 1 cm below the surface, assuming the surface were 90 cm from the source, would be $(100/91)^{-2} = 83\%$.

Beam hardening So far, I have talked in terms of mono-energetic photons. In fact, as we have noted, radiation beams from X-ray tubes and linear accelerators have a broad spectrum of energies. This would not affect the attenuation if it were not for the fact that the probability of a Compton interaction is energy-dependent, falling slowly as energy increases (as Figure 4.4 shows). Thus, the lower energy photons get attenuated somewhat more than the higher energy photons, leading to a faster attenuation of the "soft" components of the beam, and a "hardening" of the beam at depth. This effect is responsible for a less-than-exponential fall-off of dose with depth.

Scattered photons However, there is a counter-acting phenomenon, namely that the Compton interactions of the primary photons generate lower energy secondary photons. Thus, with increasing depth, there is an increasingly larger "sea" of softer photons which, in the absence of the beam-hardening just described, would lead to a more-than-exponential fall-off of dose with depth. The intensity of this sea of secondary photons depends on the size of the beam, as we shall see shortly. Figure 4.12c shows the combined effect of beam hardening and scattered radiation.

Figure 4.14 shows some practical depth–dose curves of photon beams produced from linacs.

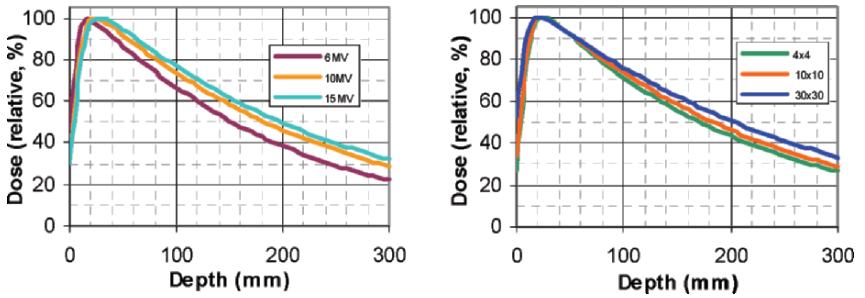


Figure 4.14. Depth dose of photon beams produced from linacs. *Left:* varying linac energies, 10·10 cm^2 field. *Right:* varying field sizes of a 10 MV linac. Data courtesy of Varian Medical Systems.

Distribution of dose laterally

The discussion to this point has dealt with the depth–dose distribution along, say, the central axis of a photon beam. It is now time to see what happens along a direction within a plane normal to the central axis dose. Figure 4.15 presents a sequence of lateral profiles, taken at some depth within the irradiated material, as we "turn on" various physical effects.

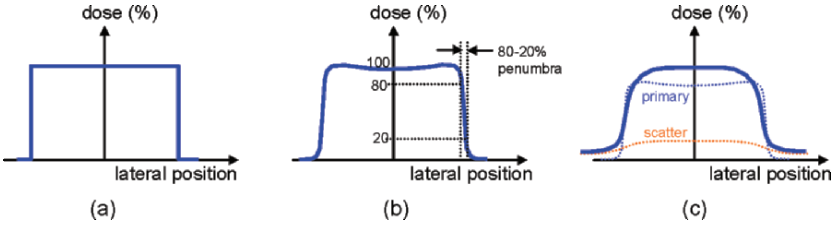


Figure 4.15. Schematic representation of the cross-field profile of a photon beam: (a) point source with perfect collimation; (b) includes finite radiation source and electron transport at the lateral beam edges; (c) includes contribution of scattered radiation (see text).

An ideal beam, with a point source, perfect collimation, and which deposits all its energy at the site of an interaction would give rise to the lateral profile shown in Figure 4.15a. This is a step-function-like dose distribution whose width is equal to the projected width of the collimators.

However, two processes cause the edge of the beam to be blurred out – creating what is called a beam *penumbra.* The first effect is purely geometric and is a consequence of the fact that the radiation source is

finite, not a point. As Figure 4.16 portrays, for points near the beam edge, the radiation source is partially, and eventually entirely, blocked by the collimator. Point P_1 of Figure 4.16, for example, "sees" all of the source; P_2 sees only half of the source and so will have ~50% of the full dose; and P_3 sees almost none of the source and so will have a very low dose. As a consequence, the beam edge is broadened by an amount which is a purely geometric mapping of the source onto the patient. For a Gaussian-shaped beam spot with a full-width at half maximum of w, the dose falls from 80% to 20% in a distance proportional to $(d_2/d_1)\cdot w$ where d_1 and d_2 are defined in Figure 4.16. Clearly, the penumbra will be smaller, the closer the collimator is to the patient.

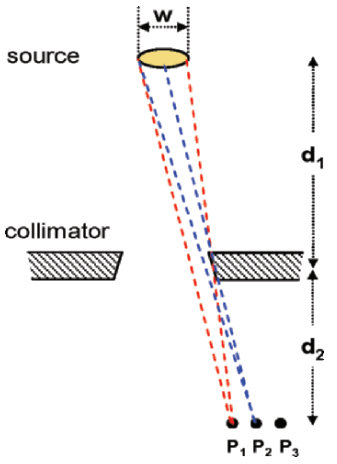


Figure 4.16. Schematic diagram showing how the dose falls off at points successively further out of the beam (see text).

The second process which blurs out the edge of the beam is that the secondary electrons created by the primary photons in the beam experience lateral dispersion due mainly to scattering of secondary electrons by atomic nuclei – a phenomenon referred to as *electron-transport*. Thus, the dose distribution does not exactly follow the intensity of the primary photons, but is blurred out somewhat, adding to the penumbra as depicted in Figure 4.15b. The penumbra grows larger at increasingly large depths, due largely to the factor d_2/d_1 in the above equation. The "cupping" of the dose is due to the influence of the flattening filter.

The last factor which affects the lateral dose profiles is the dose contribution of scattered radiation. This contribution is comprised of the secondary photons produced by Compton interactions of the primary photons and, to a much lesser extent, by further tertiary photons produced by Compton interactions of the secondary photons, and so forth. These are the photons depicted by blue dot-dash arrows in Figure 4.10, above. Their influence on the lateral profile is depicted in Figure 4.15c.

How much of the total dose is due to scattered radiation? The answer depends on three factors: the energy of the primary beam; the depth;

and, very importantly, on the size of the radiation field.[9] So far as the beam energy is concerned, the Compton interaction probability is somewhat higher for a low energy photon than for a high energy photon, and so the secondary photons produced in a high energy photon beam are more likely to escape the patient and therefore deposit less dose than the secondary photons produced in a low energy photon beam. In consequence, higher energy beams have a lesser fraction of the total dose due to scattered radiation.

So far as the changes in a beam with depth are concerned, there are two important effects. First, there are increasingly fewer primary photons at greater depths due to the beam attenuation by Compton interactions. Second, because secondary photons tend to be emitted in the forward direction, there is an increasing number of secondary photons in the beam at greater depths. These two facts together result in the fractional contribution of scattered radiation to the total dose being increasingly greater at increasingly greater depths.

The most interesting effect is the influence of the field size on the amount of scattered radiation at a given point of interest. Consider the point, P, in Figure 4.17. It lies on the central beam axis and, when irradiated by the inner smaller beam, will receive a certain dose which has primary and secondary components. Now let us widen the beam to the size of the outer beam outline. We have done nothing to disturb the impact of photons within the smaller beam. Primary photons within the stippled volume between the two beams will have no effect on the dose at P because they are

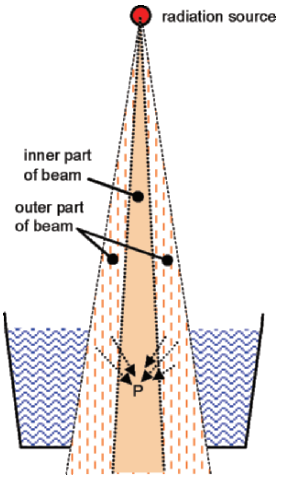


Figure 4.17. Schematic representation of the scatter dose reaching P from distant primary photons (see text).

[9] The '*field*' of a beam is, in essence, its cross-sectional shape. More precisely, its borders form the boundary of the beam's photons in a plane perpendicular to the central axis of the beam (usually at the beam isocenter). The '*isocenter*' of a gantry is the point in space about which it rotates – i.e., the center of the smallest sphere through which the central axis of the beam passes as the gantry rotates through its full angular range.

not directed towards P. But, such outer primary photons will create secondary photons – i.e., scattered radiation – and that radiation can and will reach P and will add to the dose it receives. That is, *as the field size increases, while the primary component of dose remains the same, the dose due to scattered radiation increases, and so the total dose received by P increases*. The proportion of the dose due to scattered radiation as a function of field size and depth, for 4 and 10 MeV linacs, is shown in Table 4.1.

Table 4.1. Approximate proportion of the dose due to scattered radiation on the central axis of a 4 and 10 MeV linac beam as a function of depth and field size.

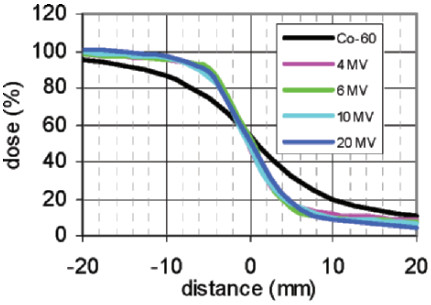
	field size		
depth	5x5 cm^2	10 x10 cm^2	20x20 cm^2
4 MeV linac			
5 cm	8%	11%	12%
10 cm	9%	16%	21%
15 cm	10%	19%	26%
10 MeV linac			
5 cm	5%	5%	5%
10 cm	5%	8%	10%
15 cm	6%	10%	13%

One sees clearly from Table 4.1 that the proportion of the total dose due to scattered radiation is: a) strongly dependent on field size; b) strongly dependent on depth; and c) strongly dependent on the beam energy. Having said this, one should also note that the magnitude of the contribution of scattered dose to the total dose is modest – ranging between 5 and 26% in Table 4.1.

The lateral distribution of the scattered radiation is broad, since the secondary photons can travel a long way (tens of centimeters). It has two effects. First, as the dose from scattered radiation rolls off slowly with distance from the central beam axis, it tends to blur out the penumbra. It reduces the dose within the field near its edge, and produces a wide-ranging low dose tail outside the field. The dose distribution *within* the field can be corrected by suitably modifying the primary photon fluence, for example by adding a suitably tailored intensity modifying device (Biggs and Shipley, 1986), but nothing can be done about the long tails *outside* the field. The scattered radiation outside the field delivers a fairly low dose relative to the

central axis dose, but it is by no means negligible. The scattered close outside the field is, for example, a source of concern regarding somatic injury to the fetus when a pregnant woman has to be treated with radiation.

Figure 4.18a shows partial lateral beam profiles for several treatment machines of different energies. The ^{60}Co therapy machine has a much wider penumbra than the linacs, due to its large source size. Note the tail of scattered radiation outside the beam at about the 10% level – consistent with Table 4.1. Figure 4.1b shows that the size of the penumbra of the linacs is relatively insensitive to energy at a given depth, but are somewhat depth sensitive, especially for the lower energy beams.



(a)

depth	*^{60}Co*	*4MV*	*20MV*
d_{max}	12	5.5	7.8
5 cm	14	6.4	7.9
10 cm	17	7.4	8.4
20 cm	24	9.6	8.8

(b)

Figure 4.18. (a) Graph of the lateral dose profile of one edge of a 10x10 cm^2 radiation field at 10 cm depth for various beam energies as indicated. (b) Table of penumbral widths (80% to 20%) measured in millimeters for 10x10 cm^2 fields of three different energies. The ^{60}Co machine is a Theratron 780. The linear accelerators are all Varian Clinacs. d_{max} is the depth of maximum buildup and is, respectively, 0.5, 1.3, and 2.5 cm for the ^{60}Co, 4 MV, and 20 MV machines. Data courtesy of S. Zefkili, Institut Curie, France.

Let us summarize where we are. We have touched on the ways individual photons and electrons interact with atoms. We then looked at how this explains the behavior of photons in bulk matter. Then, we finally discussed the generation and properties of a practical simple therapeutic photon beam. Note the word "simple." What we have not yet discussed is how to shape the beam, other than producing a rectangular field with straight-edged collimators. Nor have we touched upon the matter of varying the intensity within the field –

schematically represented by the intensity modifying device shown in Figure 4.11, but mysteriously omitted from the discussions to date. We now turn briefly to these matters.

SCULPTING A TREATMENT BEAM

So far I have addressed the design of uniform rectangular fields. Now let us see how to design a beam delivering a more sculpted dose distribution.

Beam shaping

Probably the development that spared the largest volume of tissue from unnecessary irradiation was the introduction of the technique of casting patient-specific irregularly–shaped beam blocks and apertures using Wood's metal, a low melting-point alloy (Powers *et al.*, 1973). Since the introduction of this technique, I estimate it must have spared a minimum of some 2 million liters of tissue in patients, worldwide.

Such blocks and apertures can be precisely positioned below the treatment machine head and used to block the beam where it is not wanted. Typically, such shields reduce the beam intensity to a few percent and supplement the primary collimators which provide a beam transmission of about 0.1%, but can only form a rectangular field. The design of beam–shaping apertures is addressed in Chapter 8.

A more recent development has been that of the multi-leaf collimator. This is a set of some hundred pair–wise opposing metal leaves which can individually be moved in and out by motors. Each leaf typically shadows a region some few millimeters wide at isocenter, but is thick in the direction of the beam so as to attenuate it strongly. By adjusting these leaves, the radiation field can be shaped almost at will. Used as a replacement for cast apertures, a multi-leaf collimator provides a slightly less sharp penumbra than a regular collimator, but it saves the effort of fabricating the apertures and allows beam shapes to be modified or set without entering the treatment room. However, the real power of multi-leaf collimators comes into play when they are reshaped during the course of a beam irradiation. This takes us to the topic of:

Intensity modulation of a beam

Figure 4.11 showed an "intensity modification device" placed in the beam. So far, we have discussed only so-called "open fields" in

which the intensity modifying device is not present and within which the photon intensity is relatively uniform throughout the field. There are times, however, when one wants a non uniform–intensity beam. These fall into two categories:

- Standardized *linearly varying intensity distributions.* These are used either when beams are combined (e.g., treatment using a pair of beams at 90° to one another) or to roughly compensate for a sloping patient surface. Such distributions are formed by interposing appropriately angled wedge–shaped hunks of metal into the beam (so-called "wedge filters").
- Patient-specific *intensity-modulated fields.* These can be used either for providing varying degrees of beam attenuation throughout the field to compensate for an irregularly shaped patient surface and/or internal inhomogeneities, or for intensity-modulated radiation therapy as discussed in Chapter 9. Intensity-modulated fields can, as was done before the advent of multi-leaf collimators, be made of metallic irregularly formed attenuators (looking much like that schematically depicted in cross section in Figure 4.11). However, especially for intensity-modulated radiation therapy, they are most commonly created by dynamically modifying the position of each leaf of a multi-leaf collimator and, thereby, the size and shape of the field, during the course of the delivery of a beam.

How should the shape and intensity profiles of beams be designed? Well, I am going to defer answering this question until Chapters 8 and 9. The final topic I want to briefly discuss in this chapter is:

DOSE CALCULATION

Herring and Compton (1971) presented an influential paper entitled "The degree of precision required in the radiation dose delivered in cancer radiotherapy". In this paper, they discussed dose–response data and clinical incidents in which wrong doses were delivered to groups of patients due, for example, to errors in dosimetry. They concluded that "the therapist needs a system which will permit him to deliver the desired dose distribution […] to within ±5% or possibly even more accurately." (I am sure you will share my frustration that no indication was given of the confidence level with which this accuracy was to have been obtained.) This paper, together with much other consideration of the problem, focused attention on the need to measure and calculate dose distributions accurately.

This having been said, when I entered the field in 1971, one of the main occupations of medical physicists was the calibration of therapy machines by measuring the dose to, usually, a Victoreen ionization chamber placed with a build-up cap in air at the machine's isocenter. Medical physicists placed too much emphasis on delivering the correct dose in the middle of the field rather than on whether or not the beam covered the target in its entirety or spared adjacent organs at risk at least partially. While things have changed markedly since that time, I am still amazed at how often almost the only question that medical physicists ask about a new treatment planning program is "What dose algorithm is used?" There is so much more to a treatment planning system than the dose algorithm – which generally is responsible for only a very small fraction of the software code.

For me, the calculation of dose is both extremely straightforward, and very difficult – all depending on the accuracy one requires. On the straightforward side, it is my opinion that an experienced person can, by eye, estimate the dose from a single beam to no worse than ±5% (SD). One has only to look at Figure 4.14 to appreciate that the dose fall-off is quite straightforward – about 2% per centimeter over a range of energies. The variation of dose at depth with field size will add some complexity to this, but an experienced person will have Table 4.1 in his or her head and will have no trouble factoring that in. So, assuming that this same experienced person can judge depth to ±2 cm, the goal of ±5% would seem achievable. A computer armed with even a simple algorithm could do better, and be more reliable.

On the other hand, there are all sorts of complex effects which must be taken into account for more accurate dose computation. Allowance for irregular and sloping entrance surfaces, dose perturbations in inhomogeneities (e.g., lung), loss of electronic equilibrium under several circumstances (such as when a beam traverses an air cavity such as the bronchus), allowance for edge-scattering by the collimators and photon penetration through secondary collimators, electron contamination from material in the beam – all need to be taken into consideration to achieve very good accuracy of dose estimation.

The description of practical dose computation algorithms is too large a subject for me to describe here, and I won't try. Suffice it to say that, between pencil beam algorithms, convolution/superposition algorithms and, especially, Monte Carlo calculations, very good accuracy is now obtainable.

5. BIOLOGY MATTERS

INTRODUCTION

[*]When tissues are irradiated, a complex and not fully understood chain of events takes place. At the highest level one can simply say that the radiation interacts with the tissues "through physics". Then chemistry takes over, followed by biology. This progression is illustrated in Figure 5.1.

Dose has always been the meeting ground between radiation oncologists and physicists and, indeed, between radiation oncologists with one another. The one asks for a given dose or dose distribution to be delivered to the patient; the other provides it. When asked to explain my work, I often say that we physicists are the pharmacists of

[*] It takes considerable courage even to address the subject of the biology of radiation therapy, given the availability of Eric Hall's fine textbook (Hall, 2000) – I am very happy to acknowledge its, and his, influence on me.

radiotherapy. The medication we dispense is radiation; we are responsible for ensuring that the patient is given the prescribed dose of the prescribed quality. But, and I cannot emphasize it too much, *dose is only a surrogate for what is clinically important* – to the patient and to his or her therapists. What we care about is cure and morbidity; to achieve the one and to avoid the other, as far as possible. Our goal is biological.

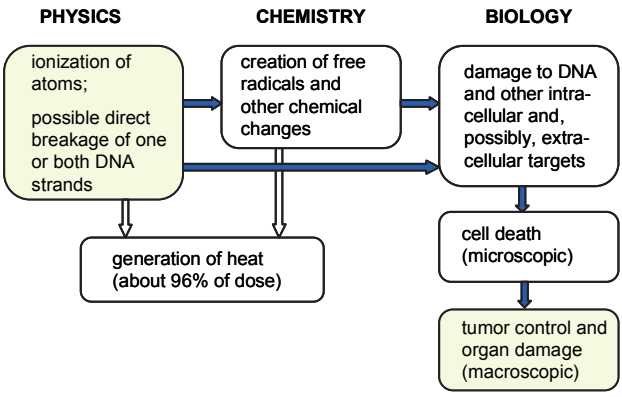


Figure 5.1. Block diagram suggesting the sequence of events following the irradiation of tissues, passing from physical effects, through chemical effects, to cellular and eventually tumor and normal tissue damage. The start and endpoints are highlighted.

In fact, it is amazing how good a surrogate dose is – especially given that so little of it goes into chemical change, as mentioned in Chapter 4. But, having said that, there are many important factors that modulate the clinical effect resulting from a given dose. To name but a few:

- the fractionation scheme (e.g., the inter-fraction interval, dose per fraction, etc.);
- the inherent radiosensitivity of the tumor and normal tissues;
- the radiation sensitizing effects of concomitant disease (e.g., diabetes), concomitant chemotherapy, genetic differences (e.g., ataxia telangiectasia), and so forth;
- the degree of oxygenation of the cells of the tumor and normal tissues;
- the way dose is distributed within the tumor and normal tissues.

As a consequence of these and related considerations, neither the radiation oncologist nor the physicist can dare to design or deliver a radiation treatment without anticipating the biological consequences for the patient of the physical deposition of dose.

MODELS

If one is to take the biological consequences for the patient of the physical deposition of dose into account, then one must have a model or models of the processes that lead from dose to damage. The development of biophysical models is held in quite some suspicion, and for very good reasons, I must say. But, the fact is that *every radiation oncologist has a number of mental models in his or her head when designing a plan of treatment*. It cannot be otherwise. How can one decide on the dose to deliver if one does not have at least a mental model of how tumor control and normal tissue damage depend on dose? Or, how can one decide how much dose can be tolerated by a particular volume of a critical normal tissue without having a mental model of normal tissue tolerance as a function of dose and volume? [1]

Those who develop biophysical models attempt to capture, in a mathematical recipe, that which is in the clinician's head and make it explicit. Once it is explicit, a few things follow: the model can be examined and discussed with colleagues; the model can be clinically tested, leading one hopes to improvements in the model; and the model can be implemented in a computer that can then allow the planner, or the computer itself, to make use of it in designing a plan of treatment.

Established experience

Radiation has been used in the treatment of cancer for over a century – incredibly, the first treatment of a cancer with radiation came only a

[1] The terms "organ" and "normal tissue" are both used in discussing radiation damage to nonmalignant tissues. In the case of the former, one usually has in mind an organized body part, such as the kidney, whose damage is expressed in terms of an impairment of the organ's function(s). In the latter case, one generally thinks of a tissue compartment, possibly with a less clear function. For the most part, the terms are used interchangeably, and I will generally use the term "normal tissue" for both meanings.

few months after Roentgen's discovery of X-rays. And, "modern" radiotherapy featuring supervoltage (≥1 MeV effective energy) radiation equipment and so-called *three-dimensional conformal radiation therapy* (3DCRT) – which features the use of multiple uniform intensity beams directed and shaped according to the patient's tumor and anatomy – has been in use for decades. This has given rise to a large body of clinical experience for a certain set of fairly well-accepted practices. I term these *established experience*.

There are a number of established treatment regimes which, for example, deliver the entire dose in only a few, or even one, fractions, or which deliver more than one fraction per day. However, established experience, as I refer to it in this book, involves the delivery of radiation in multiple (say, 30 to 40) daily fractions over several (5 to 8) weeks with a small (1.8 to 2 Gy) dose per fraction. The radiation is often delivered in two courses: the first covering suspected extensions of disease into regional lymphatics as well as the gross tumor volume, and the second concentrated on just the gross tumor volume. Beams are designed and shaped to deliver as uniform a dose as possible to the tumor and to irradiate as little uninvolved tissue as possible, and beam directions are chosen to spare sensitive uninvolved adjacent tissues, where possible.

Therapeutic ratio

Whenever one departs from established experience, both the tumor and normal tissue responses will be altered. One is little interested in either of these independently, rather one wishes to know whether, with the new regime, one does relatively better as regards its impact on the tumor as compared with its impact on normal tissues. For example, one hopes that better tumor control will be obtained for no change in morbidity or, conversely, that less morbidity is experienced for the same tumor control – or something in between. When this obtains, one says that one has improved the *therapeutic ratio*. In some situations, one can give a quantitative meaning to the therapeutic ratio, but, for the most part, the term is used qualitatively.

Types of models

There are several factors affecting radiation therapy, which one might wish to model. Among these are:

Overall time One would like to model the effect of varying the overall time of treatment. In a very general way, shorter times tend

to be harder on normal tissues, and longer times risk increased tumor proliferation.

Dose and dose-fractionation The variables here are the total dose delivered (i.e., the distribution of absolute dose), the dose per fraction, the number of fractions per day, and the number of rest days per week (usually, the two weekend days). There is, in practice, a strong correlation between these factors and overall time, since short overall treatment times tend to require the delivery of high doses per fraction and/or more than one fraction per day. The total dose that a patient can tolerate when the treatment is given in a single fraction is about a factor of three less than when eking it out over many fractions, as in established experience. Total dose is not a good surrogate for effect under all conditions! (Generally, when a statement of dose is given, the fractionation scheme should also be indicated.)

Dose–Volume effects The response to radiation of a tumor and/or of normal tissues also depends on the distribution of dose within them. In established experience, one has tended to specify the delivery of a uniform dose to the tumor, and has tended to specify a single dose in prescribing constraints on normal tissues.

Biophysical models should, in principle at least, take at least all these factors into account, for both tumors and normal tissues, so that the change in the therapeutic ratio can be assessed for any particular change in practice.

I have spent a significant portion of my life in thinking about biophysical models[2] and I want to give you my strictly personal view regarding the above three factors. Simply stated, I regarded the first two problems listed above as being too difficult for me to be able to make any useful contribution to their modeling. I have focused my

[2] I got involved in biophysical modeling through two experiences. The first was a comment by a colleague, S. Graffman, when I was worrying a lot about the possibility that, in proton therapy, inhomogeneities within the patient might lead to regions of reduced dose that might compromise tumor control. "Why don't you just try calculating how big an impact that might have?" asked my friend, who had done similar calculations himself. And so I did, and was partially comforted. The second experience was in a working group organized by the NCI in the 1980s to assess the then-new field of 3D conformal therapy. Members were agonizing about how accurate, in dose and space, dose calculational algorithms need to be. "Why don't we just try calculating the biological consequences of any errors?" I asked. And so we did. I took on TCP, believing it to be simpler, and J. Lyman was brave enough to take on the modeling of normal tissues.

efforts on understanding and modeling dose–volume effects in the belief that they were somewhat easier to assess (especially as regards dose distributions within tumors), and that they were in need of attention because their effects were largely disregarded at the time I entered the field of radiotherapy. My opinion hasn't changed and, in what follows, I focus entirely on dose–volume effects. A good review of dose–volume models can be found in York (2003), and data on partial organ irradiation can be found in Seminars in Radiation Oncology (2001).

Skepticism concerning models

There are a number of biophysical models that purport to describe dose–volume effects in tumors and normal tissues, and I will briefly address some of them below and also point to some problems with them. However, I first want to make a general point.

When I began giving talks on biophysical modeling, I noticed very different reactions from audiences consisting mainly of physicists and audiences consisting mainly of physicians. The former tended to embrace the ideas and models I presented enthusiastically; the latter were highly skeptical if not downright opposed. So I included some comments concerning the credibility of the models in my introductions. But, I made different comments to the two audiences.To physicists, I advised great caution and skepticism; when talking to clinicians, I invited them to be a bit open-minded and to consider if there wasn't at least something to the ideas I was presenting.

My point in mentioning this experience is the following. I am enormously concerned that, at the time of writing, while physicists have continued to be enthusiastic, clinicians have forgotten to be skeptical. There is too little critical thought being given to the very simple ideas which the models embody, and too little concern about accepting the implications of the models. A number of deviations from established experience have been either instigated by, or have been given support from, biophysical models. There is nothing wrong in deviating from established experience, providing it is done as part of a carefully controlled clinical trial. But widespread adoption of untested deviations is very worrying.

In my view, when models do not lead one to deviate far from established experience, one can proceed with relative confidence. When the deviations from established experience suggested by a model are substantial, one should look very long and hard at what is

proposed, and why, before adopting it – and then, only with great caution. This does not mean that established experience cannot be improved upon. Major improvements in therapy have been instigated by individuals either courageous enough, or brash enough, to try something entirely new. But there can also be disasters. *Primum non nocere* (first, do no harm).

Mechanistic vs. empirical models

There are two approaches to modeling. The first is the mechanistic approach, in which one tries to understand the basic (and, one hopes, most important) mechanisms that lead from dose deposition to tumor control or normal tissue damage and build them into the model. This generally results in a number of "free parameters," say four or five, which can be fit, it is hoped, to the existing clinical data, or can stimulate animal and/or clinical experiments to determine their values.

The second approach is empirical. One does not try to model mechanisms. Instead, one looks for a mathematical function whose shape reasonably matches the data trends. Such functions tend to have somewhat fewer parameters whose value must be established through comparisons with the existing data.

My own preference is for mechanistic models on the twofold grounds that: a) if one knows, or has a pretty good idea of, at least some of the mechanisms of damage, it seems sensible to incorporate them in the mathematics; and b) to the extent that they are true to the biology, one could hope that mechanistic models would be more reliable when extrapolating from established experience. Empiricists, on the other hand, argue both that the biology is far too complex and uncertain to be captured in a simple formula with only a few parameters, and that there are too few clinical data to fit the greater (though still small) number of parameters of mechanistic models.

It is useful to know to which category a particular model belongs.

DOSE–VOLUME MODELS FOR TUMORS

Dose–volume models for tumors attempt to predict the tumor control probability, abbreviated as TCP, under conditions of non-uniform irradiation of the tumor. Such models should, at least, be able to predict: (a) the shape and characteristics of the dose–response curve relating dose to TCP under conditions of uniform irradiation; and

(b) the shape and characteristics of the curve relating dose to the under-dosed volume of a tumor under conditions of constant TCP.

TCP and minimum dose

When I first entered the field of radiation oncology, the estimation of TCP under conditions of non-uniform irradiation was very simple. It was the conventional wisdom that the minimum tumor dose determined TCP, and that all dose delivered above that minimum dose was wasted. Indeed, worse than wasted since the excess dose (that dose which was above the minimum dose) would be responsible for unnecessary normal tissue damage outside the target volume. However, this simple understanding cannot be supported by any reasonable model, and we now think that the truth is more nuanced.

One can see why the minimum dose is not a good predictor of TCP from the following argument. Imagine a tumor being treated with 35 2 Gy fractions, 34 of them covering the entire target volume and the last one only part of it. Assume also that the dose is such that after 34 fractions with uniform coverage of the entire tumor, one would expect that, on average, one viable cell would remain. In that case, after the 34 fractions, the likelihood of the tumor re-growing if no more dose were given would be 63% – i.e., a TCP of 37%.[3] Another 2 Gy would be extremely likely to inactivate that last viable cell. However, let us assume that, for some reason, the 35th fraction only covers 90% of the tumor. The chance that the last cell will find itself in the high dose volume is then approximately 9 parts in 10. That is, there is an approximately 90% chance that the final fraction will inactivate the last cell. The TCP would then be about 37% + (63%·90%) = 94%, whereas if the TCP were determined only by the minimum dose, the last fraction would be wasted and the TCP would still be 37%.

Using a very simple expression for cell survival (assuming survival is simply an exponential function of the dose), Brahme (1984) showed elegantly that, in the event that a tumor is non-uniformly irradiated: (a) the mean (average) dose to the tumor is a good predictor of the TCP; and (b) that the TCP can only be lower than the TCP which would obtain if the tumor were irradiated uniformly to the level of the mean dose. The amount by which the TCP is actually lowered is

[3] The chance of any given number of cells surviving is determined by Poisson statistics. If the average number of surviving cells is one, then Poisson statistics tell us that the probability of there being no surviving cells, and, hence of the tumor being controlled, is $e^{-1} = 0.37$.

proportional to the second moment of the dose distribution within the tumor. This means that, the more inhomogeneous the dose distribution is, the more the TCP is lowered from the value one would estimate if the tumor received the mean dose uniformly.

TCP: mechanistic models

Brahme's model, while highly suggestive, is based on too great a simplification of cellular response to radiation for it to be used in clinical applications. There is, in practice, essentially only one mechanistic model for estimating the tumor control probability (with, of course, some minor variants). This model is based on the following assumptions:

1. Tumors consist of a large number of malignant cells, at least a fraction of which are capable of cell division and hence growth;
2. these cells do not communicate with one another – hence the fate of a given irradiated cell is independent of the fate of other surrounding tumor cells;
3. the radiosensitivity of the tumor cells within a given patient is essentially constant, but between patients radiosensitivity varies according to some distribution of sensitivities;
4. the tumor is controlled when all its cells have been made incapable of cell division.

The extent to which these assumptions hold is a matter for debate. In particular, the second assumption is certainly not rigorously true. Very interesting observations of the so-called bystander effect have made this point clear (Hall, 2003). The bystander effect is manifested in two ways. In the first, cells grown in a culture medium are exposed to lethal doses of radiation. The medium is taken from these cultures and used for a second cell culture. When these subsequent cells are irradiated, they show a greater radiation sensitivity than if they had been grown up in fresh medium – and also increased chromosomal aberrations, mutations, and oncogenic transformations. In the second manifestation, very precisely directed particle beams are directed at single cells grown in culture. When examined, the neighbors of the struck cells show increases in chromosomal aberrations, cell lethality, mutation, and oncogenic transformation.

Early TCP models did not include the third assumption; all cells of all patients were assumed to be equally sensitive. This assumption led to unreasonably steep dose–response curves. The shallow slope of the dose-response curves could be explained either if there were a very small number of cells capable of cell division – typically, only a few

hundred within a tumor of at least 10^{10} cells (Tepper, 1981) – or if there were a distribution of sensitivities among patients. As the former was judged unlikely, assumption (3) was invoked, and experimental data concerning the observed distribution of cell sensitivities supported this understanding. Recently, there is a suggestion that the number of clonogenic cells may be much smaller than was previously thought (Chen *et al*., 2006; Huff *et al*., 2006).

Under the above assumptions, one readily comes up with a mathematical prescription for estimating TCP under conditions of non-uniform irradiation. The basic approach is as follows:

1. One divides the tumor up into tumorlets, which are subvolumes (the i'th of which has a volume v_i) small enough that the dose (d_i) is essentially uniform within each one.
2. A dose-response model for the entire tumor of a given radiosensitivity, uniformly irradiated, is represented by a sigmoid curve whose slope, γ_{50},[4] and the dose needed to achieve 50% TCP, D_{50}, are based on clinical experience.
3. The dose-response for a tumorlet of volume v_i is deduced from that for the entire tumor (of volume V) through the relationship
$$\mathrm{TCP}(d_i, v_i) = \left[\mathrm{TCP}(d_i, V)\right]^{v_i/V}.$$
This relationship is a very general one, based on Poison statistics and the assumption that tumor control is obtained when no viable cells remain.
4. The TCP for the whole tumor of a given patient is taken to be the product of the TCPs for each tumorlet. i.e.,
$$\mathrm{TCP} = \mathrm{TCP}(d_1, v_1) \cdot \mathrm{TCP}(d_2, v_2) \cdot \mathrm{TCP}(d_3, v_3) \cdots = \prod_{i=1}^{n} \mathrm{TCP}(d_i, v_i)$$
5. Lastly, to get the TCP for a patient population, the individual's TCP is averaged over the presumed Gaussian distribution of the radiosensitivities of the patient population. That distribution can be arrived at by fitting the results for a uniformly irradiated tumor to the observed slope of clinical data.

---

[4] The symbol γ_P represents the slope of a dose–response curve at a level of response probability of P. It is expressed as the increase in response probability divided by percentage increase in dose. At 50% response probability, the slope is written as γ_{50}.

Details of the TCP dose–volume model(s) can be found in references (Niemierko and Goitein, 1993b; Goitein *et al.*, 1997; and Webb and Nahum, 1993). The models typically have some four to five free parameters which are determined by fits to, or analyses of, the available clinical data.

When these models were introduced, it became clear that they made some very surprising predictions, namely:

- a modest underdosage of a partial volume of a tumor might be tolerable as it might not reduce the TCP too greatly from the value it would have were it uniformly irradiated; and, on the other side of the coin,
- if one were to deliver a dose boost to a substantial fraction (but not all) of a tumor, that boost could lead to substantial gains in TCP.

These predictions had immediate clinical implications. They were the basis, in fact, for the decision at the Massachusetts General Hospital to allow, in the proton therapy of skull base tumors, a modest underdosage of the part of the tumor abutting the brain stem and/or cord in order to respect the radiation tolerance of those vital structures. The concept of allowing underdosage near a critical structure has since been widely employed, for example in radiation treatments of the prostate, where the dose to the closely abutting rectal wall has to be kept well below the dose that it is desired to give to the prostate.

So far as the consequences of underdosing part of a tumor are concerned, the TCP model described by Niemierko and Goitein, (1993b) makes prediction such as those which are illustrated in Figure 5.2. one sees that modest underdosage to a modest volume seems tolerable. For example, TCP is predicted not to be reduced by more than 3% when an underdose of 10, 7, 5, or 3% is allowed in a sub-volume of 2, 5, 10, or 20% of the target volume respectively.

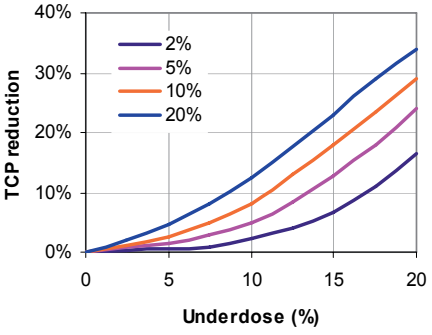


Figure 5.2. Estimates of the drop in TCP when a portion of it is under-dosed by a given dose. The curves represent different volumes of under-dosage: 2, 5, 10 and 20%.

One can similarly estimate the increase in TCP when a part of the tumor is boosted to a higher dose than the rest of the tumor. Figure 5.3 shows typical results for this situation. In this example, it is assumed that a uniform dose is given such that, if no further radiation were given, the TCP would be 50%. The model estimates the absolute increase in TCP as a function of the size of, and dose delivered by, a boost beam. The gain flattens out as the boost dose increases, but very useful increases in TCP can be obtained. For example, a 10% dose increment to 80% of the tumor volume is predicted to lead to a 12% gain in TCP, i.e., from 50% to 62%, for the particular conditions of the calculation.

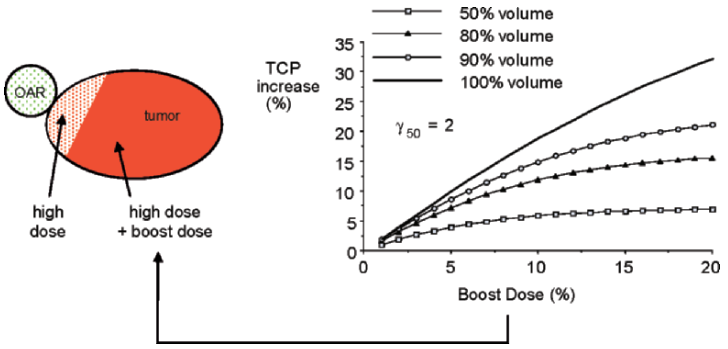


Figure 5.3. The predicted change in TCP (from a baseline of 50%) when a partial boost is given to a tumor, as a function of the volume (relative to the entire tumor volume) and dose increment of the boost.

EUD: an empirical model

Niemierko (1997, 1999) introduced the concept of *equivalent uniform dose*, or EUD, as a way of describing the impact of non-uniform irradiation of a tumor.

The equivalent uniform dose is the dose that, if applied uniformly to the whole tumor, would lead to the same TCP as would be obtained using the non-uniform dose distribution of interest. One of the appeals of this model is that it uses the units of dose. This might tend to quiet the fears of users who are suspicious of biophysical models – but, in fact, it contains, as it must, a parameter that is biological in nature. This parameter is thought of as being purely empirical, obtained by fits to the data.

For a 3D dose distribution, the EUD can be calculated as follows:

$$\mathrm{EUD} = \left(\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{d}_{\mathrm{i}}^{a} \right)^{1/a} \tag{5.1}$$

where N is the total number of voxels, and d_i is the dose in the i'th voxel and *a* is a parameter of the EUD model.

It follows that, for a differential DVH (see Chapter 6), the EUD can be calculated as:

$$\mathrm{EUD} = \left(\frac{1}{\mathrm{V}} \cdot \sum_{\mathrm{i}=1}^{\mathrm{N_b}} \mathrm{v}_{\mathrm{i}} \cdot \mathrm{d}_{\mathrm{i}}^{a} \right)^{1/a} \quad \text{where} \quad \mathrm{V} = \sum_{\mathrm{i}=1}^{\mathrm{N_b}} \mathrm{v}_{\mathrm{i}} \tag{5.2}$$

and N_b is the number of bins in the DVH, d_i is the dose in the i'th bin, v_i is the partial volume of the i'th bin, and V is the total volume.

The parameter *a* is a "biological" parameter in the sense that it is tumor or tissue specific. For tumors, *a* is generally taken to be of the order of −10. (The EUD model has also been used for normal tissues, as discussed below.)

The EUD can be used to estimate the dose "adjustment" associated with a given inhomogeneous dose distribution, relative to a reference dose of a uniform distribution, $D_{uniform}$. If one wants to translate the difference between the EUD and $D_{uniform}$ into a TCP difference, one needs only to postulate a dose–response relationship for the tumor – characterized typically by its D_{50} (the dose to achieve 50% TCP) and slope at a given TCP level, γ_{TCP}. Figure 5.4 shows schematically how an estimate of a change in TCP is made.

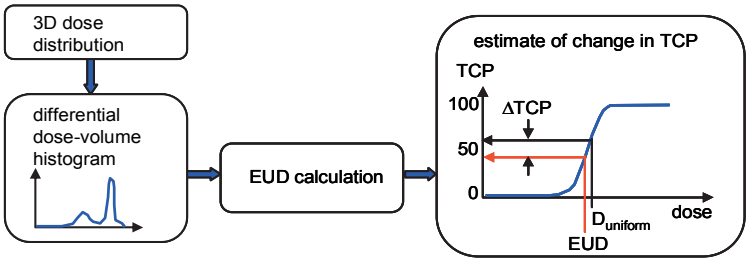


Figure 5.4. Schematic diagram of how the EUD can be used to estimate TCP. The intermediate step of computing a differential DVH can be omitted by using equation 5.1.

DOSE–VOLUME MODELS FOR NORMAL TISSUES

The modeling of normal tissue response to non-uniform irradiation is substantially more difficult than the modeling of tumor response. This difficulty stems from two main reasons. First, normal tissues are highly organized with interdependent subregions – a situation in strong contrast with the assumptions stated above for tumors. Second, while tumors are generally fairly uniformly irradiated, and so one generally does not need to evaluate the TCP for very inhomogeneous conditions, the very opposite is the case for normal tissues. So far as possible, one tends to avoid irradiating the whole volume of an organ – that is, one *wants* to apply highly non-uniform dose distributions to normal tissues. And, moreover, there is a multiple infinity of such possible non-uniform dose distributions, making comparison of the models with the meager clinical data quite problematic.

Nevertheless, it is quite clear that there are major volume dependencies in the dose–response of normal tissues and these offer clinical opportunities. The modeling of these dependencies is therefore of great potential use. One rather intriguing demonstration of the dose–volume effect is shown in Figure 5.5. In that figure is plotted the dose that is used in clinical practice for a variety of target volumes, extending from the pituitary (1 cm^3) to whole body irradiation (1,000 cm^3). These doses are, presumably, as high as possible considering the morbidity of treatment. The points in Figure 5.5 lie remarkably close to a straight line on a log–log plot and can be roughly represented, by $dose \propto (volume)^{-0.12}$. This relationship implies that, if one were to increase the irradiated volume (for example, by adding margins to allow

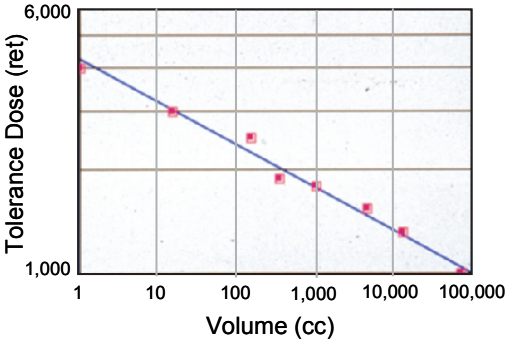


Figure 5.5. plot of "tolerance dose" as a function of irradiated volume, going from 1 cc (pituitary) to 100 liters (whole body). The dose is stated in units of rets, which supposedly corrects for fractionation differences (Ellis, 1968). This figure is based on an unpublished concept of W.S. Lowry.

for motion) then one would need to reduce the dose delivered to the target volume according to the above relationship if morbidity is to be kept the same. This relationship implies, for example, that a 25% increase in the irradiated volume would require one to lower the dose by 3% in order not to increase the morbidity of treatment.

Before describing the normal tissue complication probability models, I want to say a few words concerning the clinical "data" that have been used to fix their parameters. As part of the efforts of the NCI-funded working group, already alluded to in footnote 2 of this chapter, it became clear that, in order to be able to plan radiation therapy with the help of biophysical models, an effort was needed to ascertain the then-available data concerning the dose–volume effect in normal tissues. Based on pioneering work by Rubin and Cassarett over the previous two decades, a survey of the literature was undertaken, leading to tabulated estimates of the dose that would lead to 5% and 50% NTCPs when, one-third, two-thirds, or all of a particular organ or tissue compartment was irradiated (Emami *et al.*, 1991).[5] The then-available model of NTCP was fit to these estimates by Burman *et al.* (1991). I think it is fair to say that these papers were responsible in large part for stimulating the considerable effort to obtain more accurate and extensive clinical data that has gone on since that time.

NTCP: mechanistic models

The quantity that one wishes to evaluate in any model of normal tissue radiation response is the normal tissue complication probability, abbreviated as NTCP. When there is more than one endpoint (as there usually is), the NTCP needs to be separately evaluated for each one. The details of NTCP models are quite complex, and there are many variants. A good introduction to, and set of references for, NTCP (and TCP) models can be found in York (2003).

An NTCP model needs to be able to make predictions for completely arbitrary dose distributions. However, in working with NTCP

[5] I was a part of this effort, and I remember how difficult it was for the clinicians to arrive at these estimates for many of the 30 organs and endpoints considered. This difficulty certainly implied considerable uncertainties in the estimates, and I regret to this day that I did not heed my own words (e.g., Chapter 2) and make an effort to obtain uncertainty estimates.

models, it is common to focus attention on their predictions for what is termed "partial volume irradiation" – which is defined as the uniform irradiation of a portion of an organ, with zero dose delivered to the rest of it.

NTCP depends, of course, on a large number of factors, many of them poorly known. However, in the context of dose–volume effects, and limiting ourselves to partial volume irradiation for the moment, the NTCP is primarily considered to be a function of dose and the volume of irradiated normal tissue. It is common to use the relative volume, relative that is to the volume of the organ or normal tissue compartment. Thus, one can construct a surface in 3D that represents NTCP as a function of dose and partial volume. From the 3D representation one can draw three 2D graphs, corresponding to the three possible orthogonal cuts through the 3D surface, as suggested in Figure 5.6. It is worth bearing this decomposition in mind, because the important difference between between models is often best expressed in just one of these three graphs.

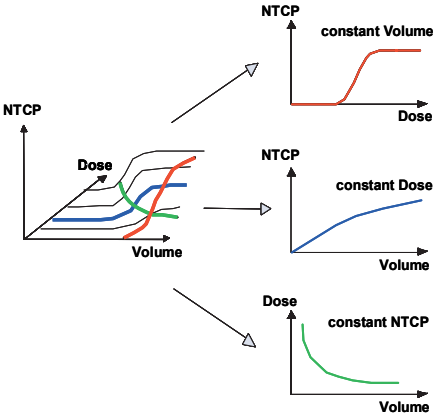


Figure 5.6. Schematic demonstration of how the 3D surface of NTCP(d,v) can be represented by three 2D graphs in each of which the third variable is held constant.

The modeling of NTCP was greatly stimulated by the publication by Withers *et al.* (1998) in which two types of tissue architecture were introduced and contrasted, namely *serial* and *parallel*. (A third tissue type with *graded response* was also defined, but it has not received the attention that the other two architectures have.) These have formed the main basis of modeling efforts since that time. Their implications are based on the hypothesis that normal tissues are comprised of elemental structures, called *function sub-units* or FSUs. Each FSU performs some function characteristic of the normal tissue, and damage of the normal tissue is a consequence of damage to its FSUs. (The nature of the damage needs to be specified, of course. Tissues may express more than one endpoint and the mechanisms need not be the same for all endpoints.) The FSU may be simply a

single cell (a stem cell, for example), or it may be a complex structure (nephrons in the kidney, for example).

Serial architecture

An organ or normal tissue compartment has a serial structure if the death or inactivation of only one of its FSUs is sufficient to cause loss of function of the tissue – i.e., to express the endpoint in question. One can imagine the prototypical serial structure as a chain that will break if any one of its links is broken – as suggested in Figure 5.7. It is thought that the spinal cord, for example, is a serial structure – although this has been questioned. The serial model is also called the critical-element model.

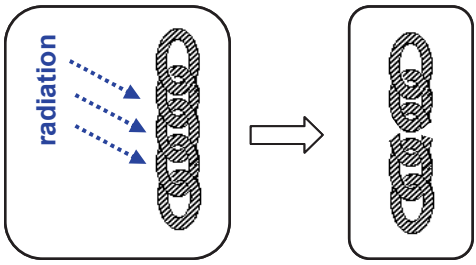


Figure 5.7. Serial architecture: normal tissue viewed as a chain which loses its functionality (i.e., load-bearing capacity) when even only one link is broken.

The central idea of mechanistic models of serial architecture is that the NTCP of the normal tissue can be expressed in terms of the probability of inactivating at least one of its constituent FSUs. The relationship would then be:

$$(1 - \mathrm{NTCP}) = \prod_{i=1}^{N_{FSU}} (1 - P(d_i)) \tag{5.3}$$

where N_{FSU} is the total number of FSUs, d_i is the dose received by the *i*'th FSU (which is assumed to be small enough that the dose is uniform throughout it), and $P(d_i)$ is the probability of inactivating an FSU that receives a dose, d_i. It will immediately strike you that equation (5.3) is very similar to the expression given above linking TCP to the TCPs of the individual tumorlets. Here, however, it is the probabilities of having no complications that are multiplicative.

Without going into the details of the models, I want to mention one point that represents the critical difference between serial and parallel architectures. Namely, *in a tissue with a serial architecture, the variation of NTCP with volume (with dose held constant) is, for small NTCPs, a linear function of volume.* This implies, for example, that if in a partial volume irradiation one doubles the volume irradiated, one will double the NTCP – provided only that it is small compared to 100%.

One can understand this striking behavior in terms of the representation of serial architecture as a chain, as sketched in Figure 5.7. Imagine that the irradiation field just covers one link of the chain, and let us assume that the dose is such that there is a 10% chance that the link will break – thus causing the chain to fail under load. In that case, the NTCP will be 10%. Now, let us double the field size so that two links are in the field, both receiving the same dose as before. Each link will have a 10% chance of being broken, so the chance that the chain will break under load (the NTCP) is roughly the sum of these probabilities, namely 10% + 10% = 20%.[6] This linear relationship is absolutely basic to serial models. It is sometimes said that "serial architecture tissues show no volume effect," by which is meant that the dose to achieve a given NTCP is essentially independent of volume. However, the argument just given shows that such a statement cannot be precisely true.

Parallel architecture

Normal tissues exhibiting a parallel architecture are also assumed to consist of FSUs, each of which performs the function that the normal tissue is responsible for. However, rather than losing function when any one FSU is lost, a parallel structure is thought to be able to maintain its function provided some critical fraction of the FSUs (e.g., 30%) maintain their function. Only when the damage to the FSUs is so great that the necessary critical fraction of them is not preserved, does the tissue itself lose functionality. The kidney and lung, for example, are thought to be parallel structures with the FSUs being, respectively, nephrons and alveoli. A parallel structure can be thought of as something like a rope, comprised of many strands, as sketched in Figure 5.8. The rope can support a load as long as a

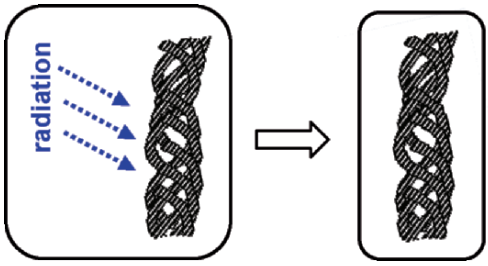


Figure 5.8. Parallel architecture: normal tissue viewed as a rope which loses its functionality (i.e., load-bearing capacity) when a critical number of its strands are broken.

[6] More accurately, following equation (5.3), we have (1-NTCP) = (1-10%)·(1-10%) = 81% . That is, the NTCP will actually be 19%.

critical number of its strands are intact. The moment that fewer strands are viable, the rope loses its functionality and will break under a specified load. The consequence of this is that, *in a tissue with a parallel architecture, the variation of NTCP with volume (with dose held constant) is not a linear function of volume but, rather, shows a threshold effect such that there is a critical volume below which the NTCP is very small, and above which it rises as the irradiated volume increases*.

EUD: an empirical model

As already mentioned, the EUD model has been extended to estimate NTCP as well as TCP. Equations (5.1) and (5.2) still hold; the only difference is in the value of the parameter, a. Normal tissues tend to have positive values of a, with serial tissues having relatively large values (say, +10 or more) and parallel structures having relatively small values (say, +0.5 to +2).

It is interesting that the simple expression for EUD, with its single parameter, can mimic both of the NTCP vs. volume (at constant dose) behaviors that have been discussed above. Figure 5.9 shows series of curves calculated using the EUD model for serial (a = 10) and parallel (a = 2) normal tissues. The linear behavior of the serial tissue and the threshold behavior of the parallel tissue are well reproduced.

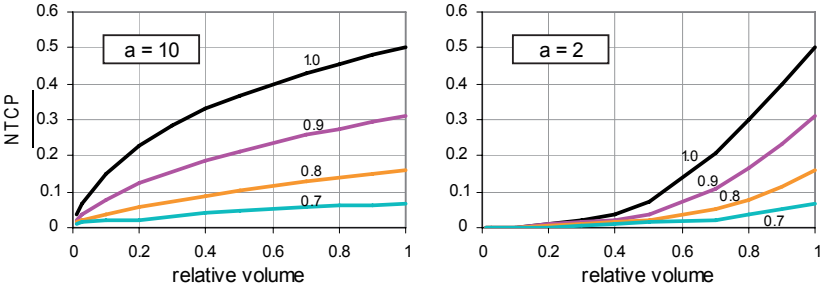


Figure 5.9. TCP *vs.* volume irradiated (relative to the volume of the entire organ) for the case of serial-like tissues (a = 10, *left*) and serial tissues (a = 2, *right*). The curves are labeled with the dose (expressed as a fraction of the ED_{50} for full organ irradiation).

The EUD has some interesting properties. For very large values of a, the EUD tends to the maximum dose in the volume. For a = 1, the EUD is equal to the mean dose in the volume. And, for very small (negative) values of a, the EUD tends to the minimum dose in the volume.

CAVEATS

But, perhaps it is not so.

Caveats concerning models of dose–volume effects of tumors

To the extent that tumors are disorganized agglomerations of non-communicating proliferating malignant cells, the assumptions in the TCP model seem reasonable. However, the model makes other, sometimes unstated, assumptions. First, that the malignant cells within a given tumor do not communicate with one another, which the bystander effect mentioned above puts into doubt. And then, that the malignant cells within a given tumor are all equally radiosensitive. It is clear that this is generally not the case. Different parts of the tumor may have quite different oxygen tensions, and oxygenation alters cell sensitivity markedly. In some tumors, the periphery may be well oxygenated as compared with the interior of the tumor and, as the minimum dose often tends to be near the tumor periphery, there is probably a correlation between dose delivered and radiosensitivity. The EUD concept has the attraction of great simplicity – but, it makes the unlikely presumption that one can say everything important about dose–volume effects with just one parameter.

The simple thought experiment presented in the section entitled "TCP and minimum dose," above, suggests pretty conclusively that the conventional wisdom that predicts that TCP is determined by the minimum dose delivered anywhere within the tumor cannot be correct. Nevertheless, one cannot say that this point is proven. For example, in Terahara *et al.* (1999), an attempt was made to correlate local recurrence of skull base chordomas with measures of dose (such as minimum and mean dose) and with EUD. Both minimum dose and EUD (but not, surprisingly, mean dose) were, in a Cox multivariate analysis, found to be good predictors of outcome! The reason for this was clearly demonstrated: because rather uniform treatment techniques were used, there was (as is often the case) a strong correlation of several of the measures – in this case between minimum tumor dose and EUD.

It seems precarious to rely on models of the dose–volume response of tumors for anything but modest deviations from established experience.

Caveats concerning models of dose–volume effects of normal tissues

Seminars in Radiation Oncology (2001) provides a review of much of the data on dose–volume effects in normal tissues, including some cautionary tales.

As stated above, normal tissues differ from tumors both in that they are internally highly organized, and that they are usually, by choice, irradiated inhomogeneously. The heterogeneity of treatment techniques is so great that it is hard to speak of established experience in connection with dose–volume effects in normal tissues. In what follows, I review a few selected experiments that illustrate the point that the models of NTCP that have been used to date cannot be the full story.

"Serial" architecture

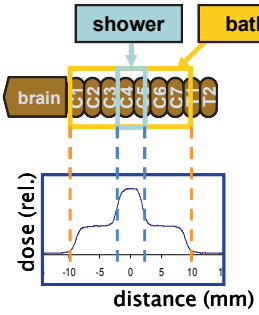
The spinal cord is the normal tissue most often cited as an example of serial architecture. In this connection, some fascinating experiments have been performed by van der Kogel, Bijl, and their colleagues on the rat. In one study (Bijl *et al.*, 2003), they irradiated first one short length, and then two separated short lengths, of the spinal cord, aiming their highly collimated beam transverse to the cord, so that its entire cross section was irradiated and the length of the irradiated section(s) was well known. They used graded doses to measure the dose needed to produce a 50% chance of leg paralysis, the ED_{50}. The dose-response was very steep, so the ED_{50} could be measured rather accurately. A sample of their results is shown in Table 5.1.

Table 5.1. Selected results from Bijl *et al*. (2003) comparing the irradiation of two separated sections with that of a single section of the rat cervical spinal cord.

	length(s) irradiated (mm)	*ED_{50} (Gy)*	*95% confidence interval (Gy)*
Single-section irradiation			
	4	53.7	49–62
	8	24.9	22–29
Two-section irradiation			
	4 + 4 (8 mm separation)	45.4	40–50

You will recall that a central tenet of serial architecture is that the complication rate is a linear function of the volume irradiated. However, in this experiment, the results strongly call that tenet into question. It is clear from inspection of Table 5.1 that the ED_{50} for irradiation of two 4 mm sections is much closer to that for irradiation of a single 4 mm section than it is to that for irradiation of a single 8 mm section. Of course, the serial architecture model would predict the opposite, namely that the ED_{50} for the irradiation of two 4 mm sections would be the same as that for irradiation of the 8 mm section since, in that case, the volume irradiated would be the same.

They undertook an even more provocative experiment, which has become known as a "bath-and-shower" experiment. In this experiment (Bijl *et al.*, 2003) a 20 mm length of the rat spinal cord (the so-called "bath") was irradiated to a sub-threshold dose,[7] and a short 4 mm length of the cord in the center of the bath field (the so-called "shower") was irradiated to graded doses and the ED_{50} of the central 4 mm section of cord (bath plus shower dose) was determined. This experiment was repeated for several different bath doses, as is shown in Figure 5.10.



Bath dose (Gy)	ED_{50} (Gy)	95% CL
0	54	49-62
4	39	37-40
12	33	32-35
18	31	26-35

Figure 5.10. The "bath-and-shower" experiment on rat spinal cord (Bijl *et al.*, 2003), showing the ED_{50} for paralysis in the shower region as a function of the bath dose (see text).

When no bath is applied, the ED_{50} for paralysis of the fore or hind limbs is 54 Gy. When a sub-threshold 18 Gy bath is added, the ED_{50} is sharply reduced to 31 Gy. Even when the bath dose is as low as 4 Gy, the ED_{50} is still greatly reduced – from 54 to 39 Gy, a 28% reduction!

[7] The ED_{50} for irradiation of a 20 cm length of cord is 20.4 Gy, and the slope of the dose-response curve is very steep, such that doses below about 18 Gy to a 20mm length of cord would not lead to paralysis.

These results strongly suggest that the radiation experience of tissues adjacent to a high dose region can markedly affect the radiation tolerance of that region. This dependency is not consistent with the serial model as formulated to date.

"Parallel" architecture

Tissues having a parallel architecture are assumed to be constituted of a large number of FSUs, all having the same radiosensitivity. However, in the lung for example, which is thought to be a good example of parallel architecture, Travis and colleagues have shown that it can be otherwise (Liao *et al.*, 1995). They developed a technique that allowed them to selectively irradiate mice, using careful collimation, delivering dose to only a third of the lung in the cephalad-caudad direction. They irradiated the top, middle and bottom thirds of the lung and assayed for both breathing rate and death from radiation pneumonitis. The radiation sensitivity of the bottom third of the lung was significantly greater than that of the apex of the lung in both assays.

There are several experiments that indicate that *the environment in the neighborhood of a region of high dose can markedly affect the response of tissues to radiation*. In what follows, I briefly summarize a few of these observations.

Lung damage as a function of heart irradiation In models of tissues with a parallel architecture, the tissue compartment or organ is assumed to react to its irradiation independently of what is going on around it. An interesting experiment (van Luijk *et al*., 2005) calls this very much into question, at least in the system these authors studied. Using careful collimation, they were able to irradiate specific regions in rats, namely: the heart, the heart plus medial lung, lateral lung on both sides, and heart plus lateral lung on both sides.

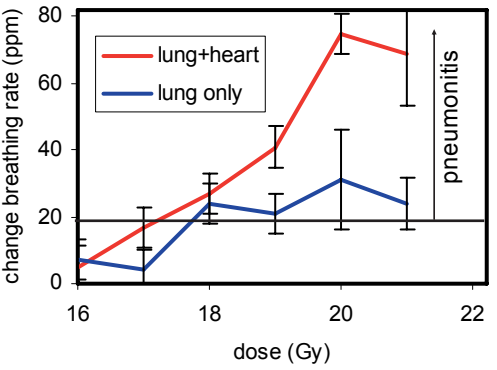


Figure 5.11. Rat breathing rate when the lateral lungs are irradiated with (red) and without (blue) the heart being irradiated. Data from van Luijk *et al.* (2005)

Using breathing rate as their endpoint, they were able to compare, *inter alia*, the response to

irradiation of the lateral lung alone, with that of the heart plus lateral lung. The result is shown in Figure 5.11. It is clear that irradiation of the heart markedly reduced the tolerance of the lung to radiation. Such a result is by no means predicted by models of parallel architecture – although one can think of physiologic reasons rather than inter-cell communication to account for these observations.

Rectal damage The geometry of "long" cylindrical and tubular organs affects their radiation response. It seems that it is the irradiated fraction of the cross-section of a cylindrical organ, or the irradiated fraction of the circumference of a tubular organ, which largely governs radiation response (see Figure 5.12). The rectum is an example of a tubular structure – and, due to the prevalence of prostate cancer and the success of high dose radiation therapy in its treatment, it is amongst the most studied examples. The rectum is very close to the prostate and so, when a high dose is given to the prostate, it is inevitable that a similarly high dose will be delivered to the anterior rectal wall. Benk *et al.* (1993) first showed, and others have since confirmed, that it is the fraction of the circumference of the rectum that is correlated with rectal complications, in this case rectal bleeding. It is now generally accepted that not more than 40% of the anterior half of the rectal wall should receive more than 70 Gy.

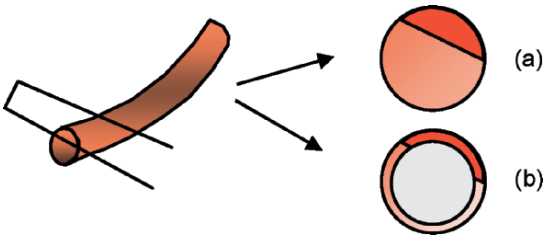


Figure 5.12. Schematic drawing of the partial-cross section irradiation of (a) a solid cylindrical, and (b) a tubular organ.

There is another suggestive finding (Jackson, 2001) that is interesting in the present context. In treating prostatic carcinoma, while delivering a very high dose to the anterior rectal wall, some patients received a higher dose to the posterior rectal wall than others. It was observed that the rate of complications was higher in the group whose posterior rectal wall received in the range of 40 to 50 Gy than in the group receiving a lower dose. That is, the radiation experience of the environment around the high dose region significantly affected the complication probability.

Paired organs The parallel models don't quite know how to deal with paired organs such as the lungs or kidneys. They can assume that the two organs act as one big organ in terms of the critical

reserve of undamaged FSUs needed to continue to function. More usually, it is assumed that the two organs are independent, and that the patient can afford to lose one of them – as is usually the case when surgery is performed. In that case, the overall probability of a complication is approximately given by the sum of the two separate complication probabilities.

However, things are not that simple in practice; the damage to the two organs may be correlated. That is, a damaged organ can cause additional damage to its irradiated partner. This was demonstrated in a study in which bilateral kidney irradiation in the mouse was compared with bilateral kidney irradiation followed 24 hours later by unilateral nephrectomy (Liao *et al.*, 1994). All assays of renal damage were less severe in the unilaterally nephrectomized group. In particular, renal tubule survival was greater in the irradiated and nephrectomized mice than in those who only were irradiated. Clearly, communication between paired organs can take place, and such communication is not taken into account by present models.

Other issues Finally, I want to make a few additional points. It deserves to be re-emphasized that any model for NTCP is restricted to a specific endpoint. The model parameters would be different for a Grade 2 complication as compared with a Grade 3 complication of some organ. There may even be different tissue architectures associated with different endpoints for the same normal tissue.

For parallel architecture organs, the mean dose is often stated to be the predictive variable of interest – for example, in the cases of the lung, liver, and parotid glands. While one must always respect the data, the limited range of techniques used in the clinical studies may not allow one to disentangle which is the truly predictive variable. Often, a number of variables are highly correlated with one another. I find it hard to imagine how the mean dose can be the fundamental variable. If we believe in the FSU concept, it would mean that a dose of, say, 20 Gy would have one third of the inactivation potential for each FSU as would a dose of 60 Gy. This would require an extremely shallow, and linear, behavior in the dose–response of individual FSUs.

And last, but by no means least, while the response of normal tissues to radiation depends on many factors, one of the most important is the fractionation scheme – i.e., the dose delivered to each point in the irradiated volume *per fraction*. Models of normal tissue complication

probability can make use of doses adjusted for fractionation effects – the so-called biologically effective dose (York, 2003) – but this adjustment certainly does not tell the whole picture.

SUMMARY

In this chapter I have tried to walk a tight line between, on the one hand, emphasizing the desirability of having explicit models of dose–volume effects in tissues and, on the other hand, citing a number of examples where the current models are inadequate. The bottom line is that such models can be useful as an aid to planning and analyzing treatments, but should never be relied upon uncritically. In particular, one should be extremely cautious when they point to approaches outside of established experience.

The examples have in part been chosen to emphasize that the environment around a high-dose region of a normal tissue can be of great importance in predicting its radiation response. Low to medium doses may appear both within an organ and in its vicinity. I am particularly concerned that we not focus only on the high-dose volume, discounting the adjacent regions of lower dose.

I entitled this chapter "Biology Matters" for a very simple reason. Namely, that the biological consequences of irradiating both the tumor and normal tissues really do matter very much so far as the success of radiation therapy, or lack thereof, is concerned. Although this is obvious, I have hardly ever heard radiobiological considerations being explicitly invoked when physicists and physicians sit down together to develop a plan of treatment for a specific patient. This may, in large part, be due to insecurity about how to convert biological issues into a prescription – and whether there are sufficient data to do so. If so, I am sympathetic, but not in agreement. The fact that something important is hard does not relieve one from the obligation to do the best one can about it. With or without explicit models, we make radiobiological judgments many times each day in radiation therapy. It behooves us to make them explicit. This will allow comparison of data and experience, and is the path to further understanding. The growth of activity in the last two decades relating to the measurement and prediction of TCP and NTCP attests to this.

6. DESIGNING A TREATMENT PLAN

INTRODUCTION

In Chapter 4, I discussed how radiation interacts with matter and, based on that, how an individual beam can be "constructed." The result could look something like the beam whose dose distribution within one section of the patient is illustrated in Figure 6.1. However,

use of this beam alone would be a perfectly hopeless way of treating the tumor. The proximal dose is higher than the already high dose delivered to the tumor and would lead to unacceptable complications in, for example, the left temporal lobe. And, the beam exits through the patient's right eye where, although the dose is lower than the tumor dose, it would still probably lead to unacceptable visual com-complications.

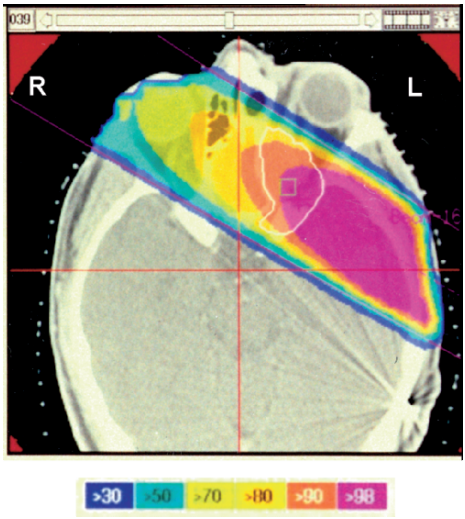


Figure 6.1. Dose distribution of a single posterior-oblique photon beam, designed to cover the target volume (white outline). Relative dose is coded by color according to the color bar below.

Thus, for all but very superficial tumors, one cannot treat the tumor using only a single photon beam. The solution is simple: to use multiple cross-firing beams that concentrate dose within the target, but spread it around outside the target so that the dose to uninvolved organs at risk (OARs) is more tolerable (see Figure 1.3 in Chapter 1). The set of cross-firing beams, together with their *weights*[1] comprise what is called the *treatment plan*. This chapter is the first of several devoted to discussing how a treatment plan is designed – an introduction to which has already been presented in Chapter 1. I have restricted my discussion to external beam therapy with photons. This restriction is for simplicity and focus. The many other forms of radiation therapy – e.g., external beam therapy with electrons, intracavitary or interstitial implants, and intraoperative radiation therapy – have, of course, additional planning issues.

[1] The "weight" of a beam is a multiplicative parameter which determines the dose it delivers. It may either determine the dose delivered to a point within the patient, or the minimum, mean or some other level of dose delivered to a volume within the patient.

THE PLANNING PROCESS

Treatment planning fits into the overall scheme of prescribing, recording, and reporting a treatment as suggested in Figure 6.2.

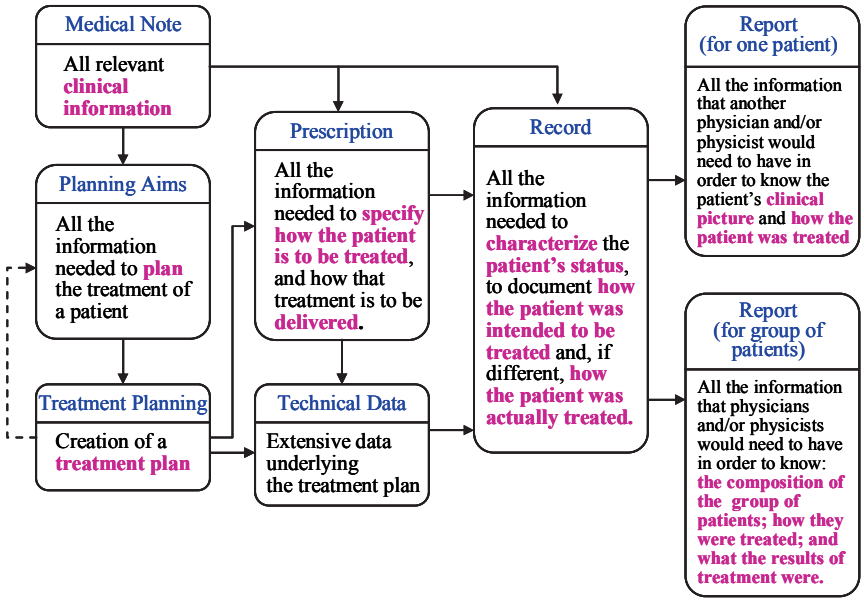


Figure 6.2. Flow diagram of the process of planning, prescribing, recording and reporting treatments. Reproduced with permission from ICRU78 (2007).

The steps of the planning process can be summarized as follows:

1. Evaluate the patient using all relevant diagnostic tools, and decide whether to employ radiation therapy as at least a part of the patient's treatment.
2. Obtain and inter-register appropriate imaging studies. This almost always includes the planning CT study that is taken with the patient lying in the position and, usually, in the immobilization device that will be used for treatment.
3. Delineate on the planning CT the target volumes (GTV, CTV and PTV) and all OARs (and, perhaps, PRVs) whose proximity

to the target volume or sensitivity makes them of particular interest.[2]

4. Establish the planning aims for the treatment.
5. Design one or more plans – i.e., sets of beams each of which, together with their weights, fulfill to the extent possible the requirements of the planning aims.
6. Evaluate these plan(s) and either select one of them for use in treatment or, if its requirements cannot be met, revise the planning aims and return to step 5.
7. Finalize the prescription.
8. Simulate the selected plan to ensure that it is deliverable and that all parameters have been correctly established.
9. Deliver the treatment, and verify that the delivery is correct, usually in many fractions over many weeks.
10. Re-evaluate the patient during the course of treatment to ensure that the plan remains appropriate (e.g., weight loss or tumor regression have not affected the treatment geometry unduly, or that there have been no unexpected toxicities) and, if it does not, return to step 5, or even 2, to replan the remainder of the treatment.
11. Document and archive the final treatment plan.
12. Review the treatment plan at the time of patient follow-up or possible recurrence.

Steps 2 and 3 have already been described in Chapter 3. Traditionally, the term *treatment planning* has tended to be used for steps 5 through 7 but I take the broader view that the task spans the whole sequence listed here and that the medical physicist or dosimetrist should be involved in all of them.

You may, with some justice, feel that step 5 is the step that, given the title of this chapter, should receive the greatest focus here. However, so much of the planning process revolves around identifying the problem(s) and evaluating the solution(s) that we need first to address these issues. The discussion of step 5 is deferred until Chapters 8 and 9 and, in the case of protons, Chapter 11.

---

[2] See Chapter 3 for an explanation of these acronyms.

PLANNING AIMS

The process of planning a treatment is almost never one of taking a prescription and translating it directly into a deliverable plan. Rather, it tends to be an iterative process, in two ways. First, the planner[3] may be dissatisfied with the first plan he or she designs and try to improve it – perhaps by using different or more beam directions, or by using a different beam energy, and so forth. Second, when despite the planner's best efforts, no satisfactory plan has been arrived at, the clinician may decide to alter the requirements – e.g., allow a higher dose to some OAR(s) – and then re-plan. This second form of iteration is embodied in Figure 6.2, where the distinction is made between the *planning aims* and the *prescription* for the patient's treatment. Planning aims are the instructions to the planner, without which he or she cannot proceed. The planning aims identify what one would like to accomplish; the prescription bows to the reality of what can practically be delivered. Of course, ideally the two will be the same, but sometimes that is not the case.

The planning aims and, subsequently, the prescription establish a number of goals. The nature of these has changed as the importance of dose–volume effects has become increasingly apparent, so that many requirements are now stated in terms of dose–volume constraints.

Requirements on the overall treatment

The clinician must specify the *prescription dose* (e.g., 72 Gy) and the *fractionation scheme* (e.g., 1.8 Gy per day, 5 days a week). As ideas have changed regarding how to prescribe treatments, the definition of the prescription dose has also changed. It used to be the desired dose at a specified point. Nowadays it is usually considered as a reference value to which the tumor dose requirement can be pegged, as discussed immediately below.

The clinician may have in mind a specific technique. For example, "our class-solution for a prostate treatment featuring a 5-beam

---

[3] Of course, there is virtually always more than one person involved in designing and evaluating a plan (physician and physicist or dosimetrist, at least) and they must reach a consensus regarding it. However, to avoid constantly repeating this caveat, I refer always to the planner, in the singular.

arrangement and using a 10 MeV linac." Or, "a beam arrangement like the one we used last month on Mrs. Jones." It would, however, not be unusual for a planner to also try to find an alternative better technique for consideration.

One special technique that the clinician might have in mind is intensity-modulated radiation therapy (IMRT). IMRT will be discussed in Chapter 9, but it is worth mentioning in the present context that, since IMRT cannot be done without computer-based judgments, additional requirements may be needed.

Requirements regarding the tumor

In essence, the tumor requirements must include a statement of both the desired dose to the target volume and the acceptable degree of dose inhomogeneity within it.

It has become common practice to prescribe the dose to the planning target volume (PTV), although it is ultimately the dose received by the clinical target volume (CTV) that matters clinically. The following are typical examples of how the dose prescription for a tumor might be stated:

- deliver the prescription dose to a specified point within the tumor e.g., isocenter or the ICRU reference point (ICRU50, 1993);
- deliver at least 95% of the prescription dose to the entire PTV;
- deliver the prescription dose to at least 95% of the PTV;

and so forth. The dose homogeneity specification could be stated as follows:

- deliver no less than 95% of the prescription dose, and no more than 107% of the prescription dose to the PTV (this was, in essence, the approach taken in prior ICRU reports);
- the minimum dose received by the PTV should be at least 70 Gy, and the maximum dose received by the PTV should be no more than 77 Gy; or
- the standard deviation of dose within the PTV should be no greater than 4% of the prescription dose.

Requirements regarding the normal tissues

The PRV is to the OAR as the PTV is to the CTV. That is, a PRV is an enlargement of an OAR to allow for motion and setup uncertainties. By rights, then, the requirements on normal tissues

should be placed on the PRVs. However, this concept has not caught on, and it is rarely the case that PRV requirements are imposed.

The requirements on OARs are almost always stated as constraints. A dose or dose–volume value is given which must not be exceeded, always with the idea that even lower values would be desirable. The following normal tissue requirements are typical, alone or in combination:

- the maximum dose to the optic disc may not exceed 50 Gy;
- no more than ⅓ of the kidney may receive more than 60 Gy and, no more than ⅔ of the kidney may receive more than 30 Gy;
- the dose to the PTV may not exceed 80 Gy.

The last of these constraints is designed to set an upper limit on the dose delivered within the PTV on the grounds that it may include normal tissue stroma whose preservation may be important to avoid long term complications.

In addition to the constraints just listed, one can set biological constraints, such as have been discussed in Chapter 5. For instance:

- the NTCP for pneumonitis of the lung should not exceed 10%;
- the NTCP for myelitis should not exceed 0.2%;
- the EUD for the bladder should not exceed 40 Gy.

There may also be constraints on the dose fractionation. The total dose allowed to be given to a specific OAR may need to be qualified, for example:

- the maximum dose per fraction delivered to the optic disc may not exceed 1.5 Gy.

Other requirements

A radiation treatment may contain two or more sequential *segments* where one segment, for example, might treat the primary tumor and regional nodes to a dose of 50 Gy, followed by a second segment in which the primary tumor is given a boost dose of 20 Gy for a total tumor dose of 70 Gy. The treatment aims and the prescription need to be separately supplied for each segment, including the *segment dose* which is, in essence, the prescription dose for the segment. Each segment of a radiation therapy course is represented by one and only one plan. The overall treatment plan, then, is the composite sum of these plans.

I have not addressed the issues of patient immobilization, target volume localization, or the management of the residual geometric and other uncertainties, which usually involve the use of treatment margins at the borders of fields. These matters, which have a strong impact on the planning process, are dealt with in Chapter 7.

Tradeoffs

It would be naïve to imagine that a treatment approach can necessarily be found which meets all the planning aims which have been set. Indeed, the ideal plan would deliver the prescribed dose uniformly to the PTV and no dose outside it. This, of course, is physically impossible to achieve, The radiation oncologist will therefore define planning aims which, based on experience, are thought to be realistically achievable. These aims will represent a tradeoff amongst the aims for the target volume, the normal tissues, and issues such as plan complexity. It is often necessary, however, to make further tradeoffs so as to arrive at a satisfactory achievable treatment plan. These tradeoffs are very much at the heart of the dosimetrist's and radiation oncologist's craft.

PRESCRIPTION

Let us assume that a satisfactory plan has been arrived at. The prescription must now be formulated and documented, and approved by the responsible clinician. What is in this prescription? In essence, it says "do that" where "that" is whatever is necessary to achieve the plan that was just approved. This means that the values of all the variables that led to the plan (e.g., the beam angles, shapes, and weights, etc.) are now to be understood as being part of the prescription. Equally, the set of planning aims that led to the accepted plan are incorporated into the prescription. These, in turn, require that the delineated volumes of interest and their underlying imaging studies need to be part of the prescription. Finally, and importantly, the calculated dose distributions and related information become at least an implicit part of the prescription. It is not uncommon to require the responsible medical physicist to sign off on these.

Not infrequently, the iterations of the planning process involve relaxation of the initially-imposed constraints. To the extent that the constraints that are finally met deviate from generally accepted values, the changes in the constraints, the reasons for the deviations, and the basis for accepting them, should be part of the record.

Technical data

In the end, an enormous amount of information is involved in fully defining a plan. Therefore, the concept of *technical data* has recently been formalized (ICRU78, 2007), as pictured in Figure 6.2. Technical data include: the planning CT scans; the delineated volumes of interest; the settings of all treatment machine parameters such as, for example, the possibly time-varying multi-leaf collimator (MLC) settings, that altogether result in the approved treatment plan; the resulting 3D or 4D dose distribution(s) and associated dose statistics; and so forth.

Seen in the larger view, the technical data are an implicit part of the treatment prescription. However, they are buried within the confines of some data management system, available only for computer recall, whereas, the prescription, almost by definition, has to be able to be written out and illustrated by sample images of the dose distribution.

REPRESENTATION OF DOSE

I am now going to make an enormous leap over the actual process of designing the treatment plan, in order to discuss how one can visualize the dose distribution that results from a given plan, and then, in the following section, how the dose distribution can be assessed. The reason for this leap, of course, is that the appreciation and evaluation of the dose distribution that results from a plan is an essential step in the treatment planning loop. One cannot discuss the design of a plan until having discussed the tools for inspecting one. The discussion of plan design *per se* is deferred to Chapters 8, 9 and 11. I will refer here to the display of dose superimposed on CT images, as CT is the most commonly used imaging modality. However, it could equally well be, for example, an MR or other imaging study.[4]

The dose distribution is part of a multidimensional data set which includes: the dose that would result from any given plan in all three spatial directions; anatomic information from one or more imaging studies; and from structure delineation – possibly including variations of these data in time. It is challenging, to say the least, to view such a

[4] At the time I entered the field of radiation oncology, dose distributions were only available as isodose contours overlaid on hand drawings of the patient's outer contour and selected internal anatomy and usually only worked out in a single patient cross-section.

data set. As a result, one normally distills the data down to lesser dimensions. In fact, one has the possibility to view 3D, 2D, 1D, and 0D (scalar) distillations of the dose distribution.[5] I discuss each of these in the following sections. However, I first want to make an important, although perhaps self-evident, point. *Whenever one abstracts information, and in particular when one reduces the dimensionality of the information so as to make it more digestible, one loses information*. Thus, one has to exercise increasing degrees of caution as the dimensionality of the information one is looking at is reduced.

Finally, we cannot ignore the fact that the view of the data that we usually have – namely a flat computer screen – is fundamentally 2D. One can add a third dimension by showing sequential 2D images in fairly rapid succession. Our eye–brain system is pretty good at fusing the sequential images so as to form a mental 3D picture. People have come up with ingenious technologies to present truly 3D images. but they have never caught on as practical tools for the inspection of dose distributions. I have tested such 3D displays myself, and have come to the conclusion that the superposition of information that occurs when viewing a semitransparent 3D world is simply too overwhelming for our sensory apparatus to accommodate. We are very lucky that the objects that populate our world are predominantly opaque, so that we see only their front surfaces, and what lies behind them is obscured. If our world were semitransparent, we would have a hard time doing something so simple as navigating our way across a furnished room.

4D dose distributions

The changes of the dose distribution in time can be short term, taking place during the delivery of a single fraction, as when the patient breathes. Or, the changes can be long term, of the order of days and

[5] Let me clarify what I mean by the "dimensionality" of a data set. A CT image, for example, is a display of intensities (representing anatomic information) in two spatial dimensions. Is this a 2D image or a 3D image? And, if I superimpose a color-wash dose distribution, do I now have a 4D image? In this book, I use dimensionality to count only the number of spatial and temporal variables. In this sense, a CT image, with or without dose color-wash, is a 2D image. But, one should be aware that the term "dimensionality" can be ambiguous, and some might include the values of the data themselves as an additional dimension.

weeks, as when the patient loses weight or the tumor shrinks. Only recently have tools such as 4DCT studies become available to follow the short-term motions of the patient. Long-term changes can be, and traditionally have been, tracked by repeating imaging studies after, say, the first weeks of therapy.

As I mentioned above, even semitransparent 3D data are simply too overwhelming to take in, and the best way we have at present of viewing time-varying dose distributions is to look at a 2D section of the patient taken at a certain time, and then sequence through the sections at progressively later times, showing a time sequence of snap-shots in a kind of movie-loop.

2D dose distributions

I am jumping over the presentation of 3D dose distributions, as 2D dose displays are the bedrock of dose viewing, and 3D presentation is largely based on 2D presentation. The solution has already been shown, for example in Figure 6.1 above. Dose and anatomic information are superimposed in a 2D image, on the screen or on paper, as here. The anatomic information (e.g., CT Hounsfield units) is represented by the image intensity at any point. Experienced users are familiar enough with normal anatomy to be able to interpret the images and identify the important normal structures as well as, sometimes, the tumor. These may be enhanced by overlaying the outlines of any delineated structures, such as the target volume as

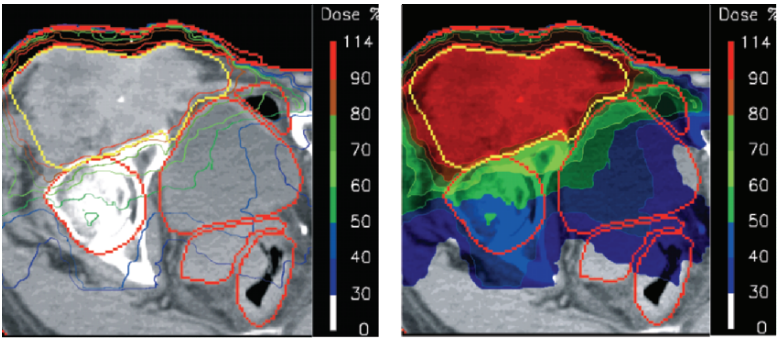


Figure 6.3. Dose superposed on a CT section with outlines of OARs and the PTV (shown as closed color-coded contours), using isodose contours (*left*) and color-wash (*right*) to represent the dose in the chosen section. Figure courtesy of A. Lomax, PSI, CH.

seen in Figure 6.1. Dose can be overlaid in one of two ways. Either, as isodose contours (lines of constant dose, much as isoelevation contours in a geographic map) the dose values of which are identified either by labels or, more usually, using colored lines. Or, as a color-wash in which the color at each point is related to the dose according to a color assignment scheme. In either case, a legend must be supplied, indicating the color-coding scheme. Figure 6.3 shows examples of these two types of display.

Of the two presentations, I much prefer color-wash displays as they are, for me, much more immediate. Their principal drawback is that the colors can flow into one another so that the boundary between two dose intervals may be unclear. This drawback can be overcome, as has in fact been done on the right hand side of Figure 6.3, by superposing isodose contours on the color-wash. Color-wash displays, however, have the disadvantage that they make it hard to appreciate fully the underlying anatomy.

The visualization of uncertainties in dose distributions (Goitein, 1985) is also, in my opinion, very worthwhile – although it is unfortunately almost never done. Figure 6.4 shows side-by-side displays of the nominal and lower- and upper-bounds on the dose at each point. One can also show the uncertainty bands on isodose contours by superposing the isodose contours of these three dose estimates.

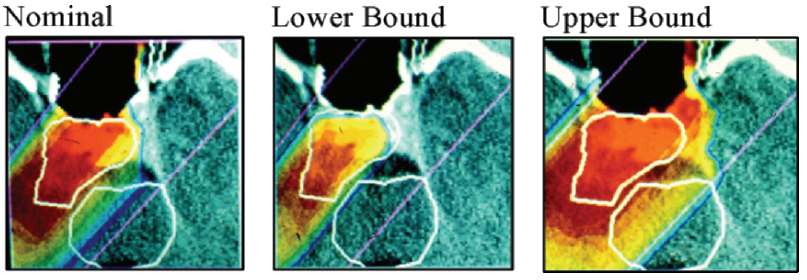


Figure 6.4. Representation of the uncertainty in dose for a single posterior-oblique proton beam. At each point within the patient's cross section, the nominal (*left panel*), lower-bound (*middle panel*) and upper-bound (*right panel*) are displayed in color-wash. One readily sees the possibility of underdose of the tumor in the lower-bound display, and of overdose of the brain stem in the upper bound display. The color scale is the same as that shown in Figure 6.1.

3D dose distributions

Returning to 3D dose representation, what are we to do? We generally can't display a 3D image *per se*. The answer has to be through the use of 2D images.

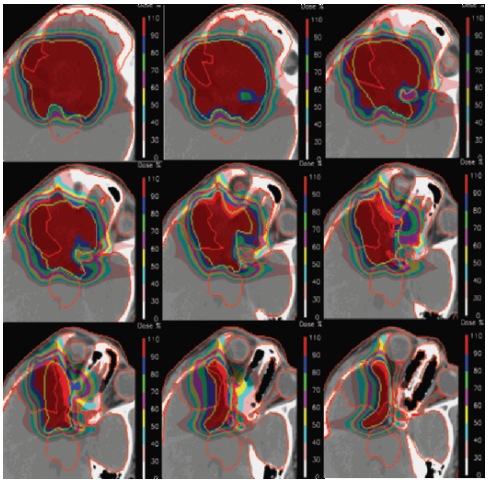


Figure 6.5. Sequence of transverse CT slices with dose superposed in color-wash. Figure courtesy of A. Lomax, PSI, CH.

The very simplest approach is simply to display a sequence of 2D CT sections all together on the screen. Such a display is shown in Figure 6.5 where, as there are only 9 images, the display is helpful. But, generally there may be from 50 to 100 sections, and then such a display is simply confusing and the images are too small to appreciate. A good approach is then to display only one image at a time, but to make it possible to scroll through the set of images in a sort of movie-loop Provided this can be done easily, under the user's control, this can be a good dose inspection method – though not susceptible to reproduction as part of the patient's documentation – or, alas, as part of this book.

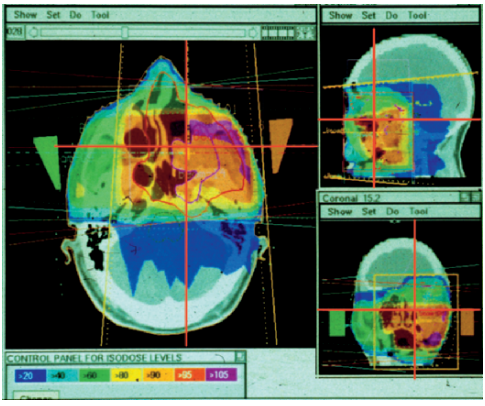


Figure 6.6. Display of three orthogonal sections through the 3D data set, with dose displayed in color-wash. In each image, the red lines indicate where the other sections intersect that image. By dragging the red lines with a mouse, the display quickly updates to show the new set of orthogonal sections.

One very good approach, now almost universally used, is to display the dose distribution (preferably as a color-wash) in three orthogonal sections; e.g., a transverse, sagittal, and coronal section. This display can be much enchanced

if: (1) the lines at which each section intersects the others are shown; shown; and (2) the user can move around the 3D space by, say, dragging the lines of intersection about and having the display update immediately. Such an display is shown in Figure 6.6.

Interactivity

The importance of being able to maneuver rapidly through these large data sets cannot be sufficiently emphasized. One needs not only a high resolution color[6] display device, but a system powerful enough to rapidly show changes in the images as one changes some controlling parameter. A refresh rate of at least 10 per second is desirable.

Time variation of dose

So far, I have described the use of sequential displays as a tool to explore the third spatial dimension. However, of course, sequentially displayed 2D images can be used to make evident the dimension of time. One would then see, in a movie loop display as described above, how the dose distribution in a given 2D section varies with time. By switching between sections, one can explore the full 4D space.

There is a subtlety regarding the representation of dose as a function of time. What one is primarily interested in is the dose at some anatomic point – let us say at the position of a particular cell – as a function of time. In general that cell will move about in time – that is, the 2D section in which the cell lies will move and distort with time. This introduces an added complication. What one wants is for the anatomy to appear to stay static and the dose display (e.g., the color-wash display or isodose lines) to vary with time. This requires that: (1) the images taken at different times be spatially registered with one another using, for example, a deformable registration technique as discussed in Chapter 3; and (2) the dose display for any particular time be mapped onto the anatomic information at some reference

---

[6] Tufte has written a series of most interesting books on the representation of quantitative data (Tufte 1990, 1997, 2001). He makes the point that everything in an image should serve a clear purpose. Colors should not be used just because they look pretty. However, in radiation oncology, the use of color is virtually essential. It is needed to code such things as anatomy (using different colors for different anatomic structures), or dose (using different colors for different dose ranges), or both.

time. Then, the viewer can see a static image with the dose moving about as time passes.

1D dose distributions: the dose–volume histogram (DVH)

The distribution of dose within a particular volume of interest (VOI) can be usefully summarized by means of a frequency distribution of the dose within the VOI – termed a dose–volume histogram (DVH) (Shipley *et al*., 1979; Chen *et al*., 1988; Drzymala *et al.*, 1991). There are two variants of DVHs: differential and cumulative DVHs. Figure 6.7 illustrates how these are constructed.

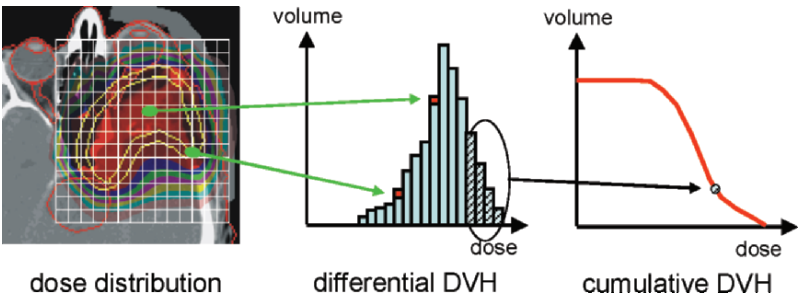


Figure 6.7. Illustration of how differential and cumulative DVHs are constructed (see text). Figure courtesy of A. Lomax, PSI, CH.

First, the volume of interest is partitioned into volume elements, called voxels, that are small enough that the dose does not vary appreciably within a voxel. The differential DVH of that VOI is the histogram, each bin of which indicates the total volume of all voxels having a dose within the dose range assigned to that bin. In Figure 6.7, two such voxels are identified. The one in the periphery of the target volume contributes to a relatively low dose bin; the one in the center of the target volume contributes to a higher dose bin.

A cumulative DVH is constructed by assigning to a given dose bin, a value equal to the sum of the volumes of all bins at that dose *and above* in the differential DVH. This is illustrated in Figure 6.7 The cumulative DVH is, strictly speaking, not a histogram at all, but a cumulative frequency distribution. Nevertheless, the nomenclature "dose–volume histogram" has become sanctified through usage. The interpretation of the ordinate of a point on the cumulative DVH is that it is the total volume of the VOI that receives a dose greater or equal to the dose indicated on the abscissa.

DVHs have become widely adopted as a tool for dose summarization – particularly when plans must be compared, as discussed below. However, they share the problem of all data abstractions in that information is lost. In the case of a DVH, one loses all spatial information about the dose within the VOI whose dose it summarizes. One cannot tell, for example, whether low doses in the DVH come from one subvolume of the VOI, or are distributed across many subvolumes. Moreover, particularly with large VOIs, the sheer volume of tissue may make it easy to miss small hot or cold spots. For all these reasons, I judge it very unwise to rely on DVHs alone to analyze a dose distribution; *DVHs should always be looked at in conjunction with graphical representations of the dose distributions*.

As between differential and cumulative DVHs, while the former have their value, they are in practice little used. A major reason for preferring cumulative DVHs is that a number of useful dose statistics can be directly read off them, as indicated in Figure 6.8.

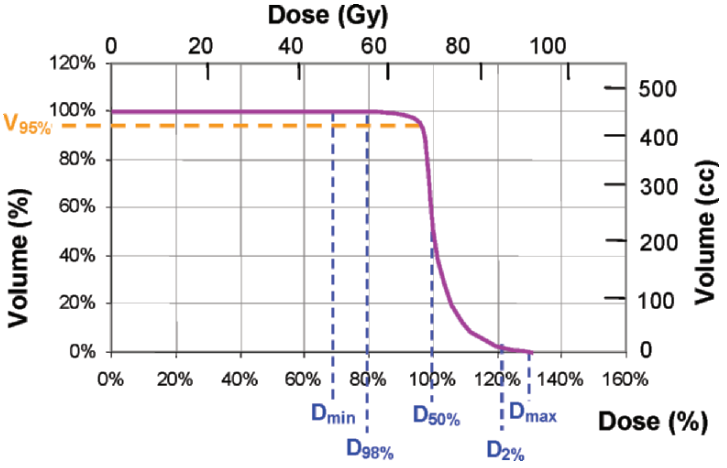


Figure 6.8. Demonstration of how a number of dose statistics can be read off a cumulative DVH. Reproduced with permission from ICRU78 (2007).

0D dose and dose–volume statistics

The "0" in 0D indicates that dose statistics are scalar quantities; they have magnitude but not direction. In describing dose–volume relationships, the following nomenclature has been established, for example, in ICRU78 (2007):

V_D is the largest volume of a VOI that receives more than or equal to the dose, D. Both the volume and the dose may be in absolute or relative units. Which is intended should be made clear by the addition of the appropriate units.[7] For example:

- V_{70Gy} = 142 ml means that "142 ml of the VOI receives at least 70 Gy"; whereas
- V_{70Gy} = 80% means that "80% of the VOI receives at least 70 Gy." Similarly,
- $V_{90\%}$ = 142 ml means that "142 ml of the VOI receives at least 90% of the reference dose."

For relative volumes, the reference volume should be identified. Usually it will be the entire volume of the VOI – either as imaged or, if not fully imaged, as estimated.

D_V is the dose that a volume, V, of a VOI reaches or exceeds. Both the dose and the volume may be in absolute or relative units. Which is intended should be made clear by the addition of the appropriate unit to numbers. For example:

- D_{142ml} = 70 Gy means that "at least 70 Gy is delivered to 142 ml of the VOI"
- $D_{80\%}$ = 70 Gy means that "at least 70 Gy is delivered to 80% of the VOI"
- D_{142ml} = 90% means that "at least 90% of the reference dose is delivered to 142 ml of the VOI."

For characterizing dose distributions, useful dose and dose–volume statistics include: the dose at a specified point; the minimum (D_{min}), near-minimum ($D_{98\%}$ or $D_{near\text{-}min}$), median ($D_{50\%}$), mean (D_{mean}), near-maximum ($D_{2\%}$ or $D_{near\text{-}max}$), and maximum dose (D_{max}) within a specified VOI; and the volume of a specified VOI receiving at least a specified dose (V_D).

(The reason that $D_{98\%}$ (rather than D_{min}) and $D_{2\%}$ (rather than D_{max}) are of interest is that, in some computer programs, errors in the calculations and in the digital representation of the VOI delineation

[7] Unfortunately, the units are often omitted when V_D and D_V are stated, and they have to be interpreted according to the presumed usage. For example, V_{20} is usually intended to represent the volume receiving ≥20 Gy, i.e., $V_{20\ Gy}$.

give rise to very small artifacts that are purely calculational and have no clinical implications.)

As just alluded to, one great advantage of cumulative DVHs is that many dose statistics – such as the minimum, near-minimum, median, near-maximum, and maximum dose for the VOI represented by the DVH – can be directly read off them, as Figure 6.8 indicates. This is not the case with the mean dose, D_{mean}, which has to be calculated.

0D measures of biological effect

Another group of scalar quantities that can be used to characterize a dose distribution are estimates from biophysical models, as described in Chapter 5. These would include: TCP and EUD for the tumor, and NTCP and EUD for specified OARs.

PLAN ASSESSMENT – THE BALANCING ACT

There are two approaches to plan assessment: (1) inspection of the dose distribution and quantities derived from it by an expert; or (2) the computation of a "score" for the plan. The latter approach is generally restricted to a computational search for the "optimal" IMRT plan and this aspect of plan evaluation is deferred until Chapter 9. However, even when a computer has arrived at a plan with the best computed score, the radiation oncologist will need to review it, and will use the techniques of *expert inspection* for that purpose.

It has already been mentioned, but it bears repeating, that the choice of a good treatment plan involves a balancing act between, on the one hand, the likely effect of the proposed irradiation on the tumor and, on the other hand, its likely effect on the normal tissues. Whether one judges a plan using quantitative biological models explicitly or by inspection from the dose distribution, the ultimate evaluation relates to the need to achieve a balance between local control and morbidity.

The planner will also have in mind the feasibility of safely delivering a particular plan in practice. This judgment is an important part of the planning task and it is one that requires a good deal of experience to make.

Organ by organ inspection

When judging plans "manually" *clinicians tend to look at the dose distributions within the tumor and within individual organs and*

tissues separately, one by one. That is, they analyze each compartment of concern one at a time. In the tumor, they look to see if the coverage, level and homogeneity of dose distribution are satisfactory – e.g., without undesirable hot or cold spots; in the OAR's they look to see if the dose they would receive would be "tolerable." Then, in some fashion or other, the nature of which is critical but very hard to analyze, they synthesize these judgments so as to arrive at an overall assessment. This assessment allows them, for example, to rank two plans so as to be able to say which of the two they prefer.

Tumor control

In assessing a given plan, the planner will want to evaluate the absolute dose and dose inhomogeneity in the target volume(s) in order to make a judgment about the tumor control probability (TCP). One helpful measure of dose inhomogeneity is the difference between the EUD for the dose distribution and, say, the mean dose. The difference is, in essence, the "lost dose" due to dose inhomogeneity.

Local tumor control can, of course, be undermined by the presence of uncontrolled metastatic disease. Important though this is, radiation therapy planning generally takes the question of metastases into account only at the strategic level of choice of modality – e.g., by employing chemotherapy in combination with radiation. Such a decision will, in turn, affect the radiation dose that can be given – expressed as a need to lower the normal tissue dose constraints, and/or the prescription dose. These influences are likely to affect the TCP which can be achieved and/or the likely morbidity.

The missing tissues

The delineation of the patient's organs and tissues is, at least at present, a time-consuming and complex process. For this reason, usually only a few such compartments are explicitly defined, leaving a large volume of tissue unaccounted for. It would not be unusual to find, in practice, that only a quarter or less of the tissue volume which may potentially be irradiated has been explicitly delineated. However, even though the planner chooses to ignore these tissues, the radiation does not! *It is important to take into account such tissues in assessing a treatment plan*. At the very least, one should evaluate the "remaining volume at risk" (RVR), which is the volume that is within the patient but outside the target volume and all delineated organs and tissues. The important task of tracking what is done to the

non-delineated tissues is acknowledged as necessary by many, but done in practice by rather few.

Combining all the factors

After having assessed the impact of the plan on the target volume, the OARs, and the RVR, the radiation oncologist will need to make an overall judgment about the plan's merit and/or acceptability, based on all of these. This judgment can be excruciatingly difficult to make. It is also very hard to analyze how experts arrive at this synthesis. My experience leads me to think that the way it most often works is that a plan will be judged satisfactory if: (1) the risk of morbidity of each OAR is acceptably low (this judgment could be reinforced if the calculated NTCP is within the constraint or constraints initially given by the clinician); (2) the likelihood of tumor control (which could be reinforced by the calculation of the TCP) is judged to be as high as possible, given the OAR constraints; and (3) there are no substantial hot spots in the OARs, or tepid spots in the target volume.

PLAN COMPARISON

For many purposes, one needs to be able to compare two or more plans with one another in order to decide what the salient differences are and, perhaps, which is "better" or the "best." Naturally, the pre-requisite for being able to compare plans is the ability to evaluate the plans individually, as has just been discussed.

Plans can be compared, just as in plan assessment, either by expert inspection, or by computing scores for each plan under consideration, favoring the plan which has the highest score. The latter approach is usually confined to the search for an acceptable computer-based IMRT plan, but it is equally appropriate to uniform beam radiation therapy. The following approaches have to do with expert inspection.

Side-by-side dose distributions

One good way to compare two or more dose distributions is to display them side-by-side in separate panels. Each dose distribution is presented in its own panel as a color-wash or isodose lines super-imposed on a CT or other image – the same image being used in each panel. The observer can then interactively page through all sections of the study, observing the dose distributions of all plans being compared, level by level.

The ability to display two or more plans side-by-side, and to superimpose DVH's from two or more plans, is a relatively recent treatment planning capability. It used to be that one had to print out isodose plots superimposed upon simple outlines of the anatomy and literally lay them on a table, side-by-side. I remember my excitement when, having conceived of the idea of side-by-side display on the computer, I was able to show for the first time doses for two plans, displayed in color-wash over the CT scan, simultaneously. This may seem like a trivial improvement, and it is now routine. But at that time it was a revelation; I could suddenly "see" the differences. The computer display was so much more immediate. In addition, one could interactively place a cursor at a point in a CT image and numerically display the doses from the two plans at that point. Figure 6.9 shows such a pair of color-wash images. In this figure, for simplicity (and, it must be admitted, to make a subliminal point) each "plan" features just a single posterior oblique beam, of protons in one plan and of photons in the other.

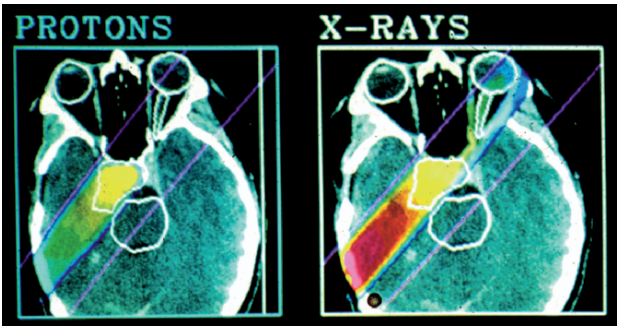


Figure 6.9. Side-by-side comparison of two (single-beam) plans with the dose distributions shown in color-wash.

But… there are dangers. The scale for the color-wash representation of dose is at the user's discretion and, either due to lack of thought or something less benign, the scale may easily conceal or emphasize selected dose regions. The colors cover a selected range of doses. Doses above the highest dose in the range are represented by an additional color which signals that the dose is higher, but cannot indicate how much higher. Regions having doses below the low end of the range are usually represented in grey scale, so that lower doses are, in effect, suppressed. Figure 6.10 shows what can happen in practice. (The figure uses admittedly a rather unusual color palette,

but the point to be made is independent of the color palette employed.) In panels (a), (b), and (c), the lower end of the color scale has been chosen to be 25, 10 and 45 Gy respectively. All three panels show the same pair of dose distributions, one of a photon plan, the other of a proton plan. Panel (a) is probably a reasonable comparison of the two plans; panel (b) makes photons look quite a bit worse than protons; and panel (c) makes protons and photons look virtually identical. Quite different conclusions could be drawn about the comparative merits of the two modalities, depending on the range of doses included in the color-wash.

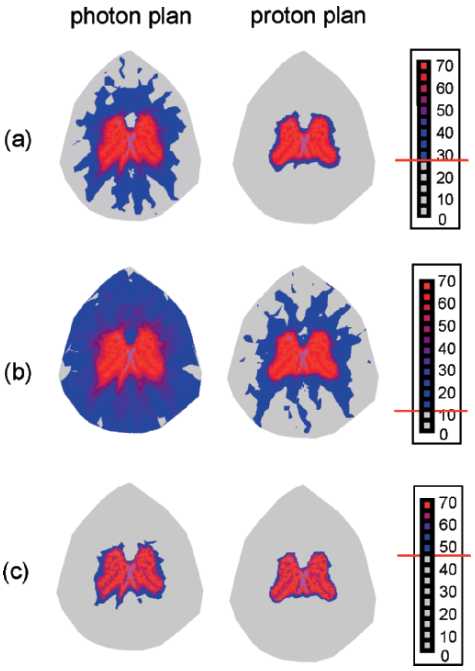


Figure 6.10. A comparison of two plans using three different low-end cutoffs for the color display (see text). Figure courtesy of A. Niemierko, MGH, USA.

This example highlights two important points. First, dose displays can be misleading. This is equally true when isodose contours are used. However, the greater immediacy of color-wash makes the perils rather starker. One must not be beguiled by pretty color pictures. Both for the reason being brought out here, and because there are always uncertainties in the computed dose distributions, one must be aware that *it is not necessarily the case that "what you see is what you get."* The ideal is WYSIWYG; reality sometimes falls far behind.

The second point is that, whenever one is looking at a treatment plan whose doses are displayed over a restricted range, one must interactively adjust the endpoints of the range to ensure that there are no hidden surprises. I already pointed this out in Chapter 3 as it regards the window and level with which CT and other images are viewed; it is equally true for dose displays.

Dose difference display

Another interesting way of comparing a pair of plans is to display the *difference* in dose for the two plans, overlaid on a CT image and using a color scale that permits the coding of both positive and negative differences. In such a display it helps greatly if small differences that are not considered to be clinically important are suppressed (i.e., not shown in color). When analyzing such an image, the size and location of the dose differences can be readily seen – whereas, when comparing DVHs as discussed immediately below, the *location* of any differences is not observable. However, the information as to the absolute dose level is lost. A 10 Gy dose difference between 75 and 85 Gy has an entirely different significance than the same difference between 5 and 15 Gy. For this reason, one should always inspect dose-difference displays in conjunction with a display of the dose distribution of at least one of the plans being compared. Such a difference display is shown in Figure 11.6 of Chapter 11.

Overlaid DVHs

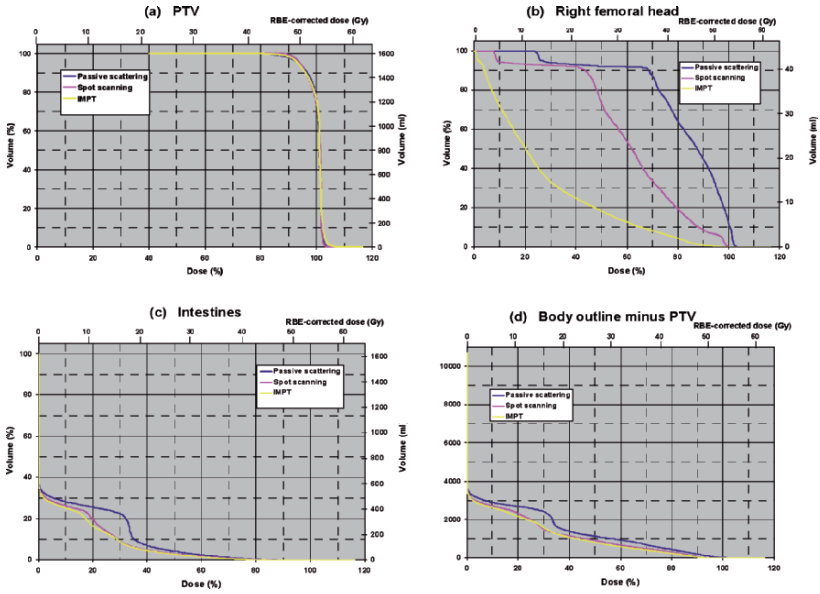


Figure 6.11: DVHs for several VOIs, with the curves for the three plans overlaid. Figure courtesy of G. Goitein, PSI, CH.

An often-used way of comparing plans is to display a number of panels, one for each volume of interest, and in each panel show the DVHs for the plans being compared. Figure 6.11 portrays DVHs for

several VOIs in which the DVH curves for three plans which are being compared are shown. (The three plans, in fact, feature the three different proton beam delivery techniques illustrated in Figure 11.13 of Chapter 11.) The DVHs are very valuable in helping one assess the dosimetric differences between the plans and, hence, between the techniques.

Commonly, two or more DVH curves cross one another – although this is not the case in Figure 6.11. One then needs to know which of two crossing curves is better, and by how much. This matter is touched upon in Chapter 8.

Comparison of dose statistics and biophysical models

Table 6.1. Side-by-side, dose statistics for the same three plans as are illustrated in Figure 11.13. Table courtesy of G. Goitein, PSI, CH.

	scattered protons	scanned protons	IMPT
Target volume (PTV)			
dose to 98% of volume, $D_{98\%}$	49	50	49
median dose, $D_{50\%}$	55	55	55
dose to 2% of volume, $D_{2\%}$	56	55	56
relative volume receiving 95% of the prescribed dose, $V_{95\%}$	93	93	92
Right femoral head			
relative volume receiving 20% of the prescribed dose, $V_{20\%}$	54	50	27
relative volume receiving 50% of the prescribed dose, $V_{50\%}$	50	40	10
relative volume receiving 80% of the prescribed dose, $V_{80\%}$	35	10	2.3
dose to 2% of volume, $D_{2\%}$ (near-maximum dose)	102	98	85
Intestines			
…			
Body outline minus PTV			
mean dose outside the PTV, D_{mean}	7.2	5.9	5.6

To compare dose statistics for two or more plans, it is very helpful to lay out the information so that the data for the plans lie side-by-side. An example of such a table is shown in Table 6.1. It is extremely helpful in such a table to include in an additional column the constraints imposed by the prescription, so that one can judge how well they have been met, and whether one plan does a better job than

another in that respect. One can also include in such a table, of course, the values predicted by any biophysical models of interest.

Combining all the factors

Just as for plan assessment, but even more so for plan comparison, the job of pulling together the many disparate data into an overall judgment is an excruciatingly difficult task. I have no guidance to offer concerning it.

POST-PLANNING ACTIVITIES

Finally, just a very few words about the activities that take place after a satisfactory treatment plan has been arrived at. These include the following.

Simulation of treatment

A *simulator* is a machine that simulates the features and geometry of a treatment machine, but substitutes a diagnostic X-ray tube for the therapeutic source of photons. Simulators were introduced as an aid to designing a treatment plan – and for efficiency, so as to avoid tying up a valuable treatment machine (Karzmark and Rust, 1972). In days of yore, the plan was often designed on the simulator by adjusting field sizes and shapes until the target coverage and normal tissue avoidance of each beam was deemed satisfactory as seen on radiographs and/or fluoroscopy.

The role of the simulator has vastly changed. Modern treatment planning systems act as "virtual simulators," allowing plans to be developed after the patient has gone home, leaving only his or her planning CT study and other images behind. However, the treatment planning systems only partially represent reality, especially so far as both patient and treatment machine geometries are concerned. Once a plan has been arrived at, it is usually tried out on a simulator or for complex equipment on the treatment machine itself, to see if there are any unanticipated interferences between the patient and the treatment equipment and to make sure that the geometry embodied in the CT and other studies remains representative of the patient.

Delivery of treatment

There is many a slip 'twixt cup and lip. The flawless transfer of consistent data between prescription, treatment plan and actual treatment is critical and is a common source of problems. When I

first started to work as a medical physicist, due to lack of space, I was assigned as my "office" a counter top in one corner of the control room of one of the treatment machines. This turned out to be an invaluable experience, giving me the chance to observe the routine practice of radiotherapy on a daily basis. I recall seeing on two occasions the radiation therapist realizing that she had just treated a patient with another patient's treatment parameters. This was the time in which computer programs to perform the so-called record and verify function (having the computer monitor the machine settings and not allow treatment to proceed unless they matched the prescribed parameters within a defined tolerance) were just being introduced. I recall the amazed and concerned reaction to the study by Chung-Bin and his colleagues (Kartha *et al*., 1975) who, using the computer as a silent monitor of treatments, found that the mis-setting of treatment parameters occurred at an approximately 3% rate and that more than two-thirds of the patients monitored had at least one error at some stage during the full course of their treatment. This ushered in the use use of record and verify systems. However, many "old hands" (and probably many new ones) have reservations about these systems, too (Klein *et al*., 2005). They assure that what is done is what the computer data base says should be done. But, this also provides an opportunity to do the wrong thing consistently, every time.

The more we become mechanized, the greater is the need for human oversight and the exercise of "common sense" – a quality that has not yet become one of the computer's skills. In my view, it is essential that both the clinician and the physicist who planned the treatment attend the first treatment and periodically thereafter, to help ensure that what was planned is what is being delivered. The importance of quality assurance is underlined in Chapter 12. To an extent, quality is supported by instrumentation of various sorts. However, nothing can replace the eyes and brains of the experts – radiation therapists, dosimetrists, physicists, and radiation oncologists – continually monitoring what is done in practice.

Ongoing patient evaluation

As is mentioned in Chapter 7, an important source of uncertainty is the possibility of unappreciated changes in the patient's condition and geometry during the generally several weeks of therapy. Thus periodic checks are generally necessary, the nature of which depend on the clinical situation. These could extend to periodic rescanning of

the patient and, if indicated, re-planning of the remainder of the treatment.

Documentation and archiving

People tend to underestimate the importance of documentation. I learnt a lesson in this regard during my doctoral thesis work. I was participating in a high-energy physics experiment designed to measure the internal structure of protons and neutrons. One of the members of our group was a Nobel prize winner. Whenever he participated in a shift during the running of the experiment, he commandeered the data books and sat at a small table in the middle of the room, meticulously recording everything that was going on around him. As a brash youngster, I could not understand why such a senior scientist would undertake what I imagined to be a secretarial job. Only slowly, over years of subsequent data analysis, did I come to appreciate the importance and value of what he had been doing. Given the selectivity and imperfection of our memory, and the need to transmit to others what we have done, it almost seems that what is not recorded, never happened – except, of course, in the case of an irradiated patient whose tissues will remember what happened to them, even if we don't.

It is essential that all data concerning a radiation treatment be recorded. It is needed in order to know, at the time of follow up, what was done to the particular patient. In the unfortunate case of a treatment complication or tumor recurrence, the information is critical for planning a retreatment or an alternative method of management. And, the recorded data are necessary for the analysis of groups of patients, so that we can learn from what we have done and thus benefit those yet to come.

7. MOTION MANAGEMENT

MOTION OF, AND WITHIN, THE PATIENT

One of the daunting problems in radiation therapy is to accurately aim a beam of invisible radiation at an invisible moving target – without missing it, and without making the beam so generous that, while it covers the target, it irradiates too large a volume of normal tissue to be acceptable.

Incredibly, when I entered the field of radiation therapy, I was under the misapprehension that I was dealing with a static problem. That my job was to understand what happened when a beam of radiation

was directed at a complex but stationary object. It took me a little while to appreciate that the patient is a living, breathing, moving individual. That I was facing, in fact, a dynamic problem and that I had better understand the time-varying processes if I were to make any headway in my new profession.

When one speaks of motion, it should be with a broad understanding of the term. As I use it, the term motion covers both short term and long term changes over time of the patient's location, the location of the patient's tumor and organs, the size and shape of the tumor and normal tissues, including uncertainties in their delineation as discussed in Chapter 3, the degree of filling of organs such as the bladder, intestines, and rectum, and so forth. All of these phenomena will affect the dose distribution due to a given beam, and all must therefore be taken into account.

To deal with motion, as with any source of uncertainty, one needs to proceed in three steps:

1. understand qualitatively and quantitatively the nature and degree of the problem(s);
2. instigate measures to reduce the size of the problem(s) to the extent that is feasible and practicable; and
3. develop and implement strategies to take the inevitable residual uncertainties into account in an optimal manner.

Tumor and organ motion can be classified into three categories, namely: motion of the patient as a whole relative to some reference object such as the couch top; intra-fraction motion of the tumor and organs within the patient (i.e., during delivery of a single fraction); and inter-fraction changes of the position and/or size and shape of the tumor and organs within the patient.

In managing a patient, taking motion into account, the sequence of procedures shown in Figure 7.1 is generally followed.

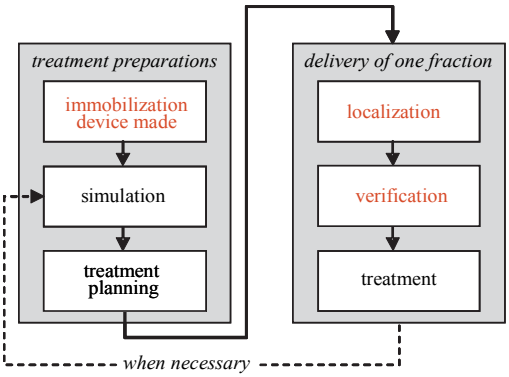


Figure 7.1. Block diagram of the steps involved in preparing a patient's treatment. The colored steps are those which are discussed in this chapter.

IMMOBILIZATION

It is common to employ some method of immobilization to better relate the patient to the treatment equipment. In some special cases, the immobilization device is built into the equipment. Usually, a separate device is used and placed upon the couch top or, much less commonly, treatment chair, often being indexed through the use of locating pins.

An immobilization device is used to hold the patient as a whole in a stable and near-motionless position during both imaging and treatment. By doing so, the locations of internal organs and the tumor are also constrained. The use of immobilization devices at the time of acquiring the planning CT serves to minimize the problem of the patient's position during treatment being different from the patient's position during the imaging studies.

The two-joint rule

One might think that all that is needed is to immobilize the body part within which the tumor lies. However, the adjacent body parts usually have an influence on the part to be immobilized and themselves need to be immobilized. For example, in treating prostate cancer, the positions of the upper and lower legs, and their degree of rotation are important in achieving reproducible positioning. A good rule (attributed to Verhey) is that body parts that are at least two "joints" away from the part within which the target volume lies need to be immobilized. This rule is illustrated schematically in Figure 7.2.

Figure 7.2. Schematic illustration of the two-joint rule. The target volume is within the cranium, but both the neck and torso should be immobilized (the blue support in this figure), as well as the head.

Immobilization techniques

A review of immobilization methods in radiation oncology can be found in Verhey and Bentel (1999). Many types of immobilization device are available, including bite-block/head-rest combinations for stabilizing the head, partial body casts for stabilizing the thorax or

pelvis, and whole body casts. Casts may be made, *inter alia*: from plaster of Paris, using conventional moulage techniques; from thermoplastic sheets that are draped over the patient while warm and become firm upon cooling; and from bags of foam pellets, that are made rigid by being placed under vacuum once the bag is made to conform to the patient's surface.

Some of the more common immobilization techniques are the following.

Perforated thermoplastic masks

A perforated thermoplastic sheet is formed to the patient's head while warm and allowed to set by cooling. The perforations keep the head cool, are less claustrophobic for the patient, and preserve some degree of skin sparing. The sheet is captured in a tennis racket-shaped frame that is attached to the couch top or chair with the use of indexing pins (Verhey *et al.*, 1982) – as illustrated in Figure 7.3.

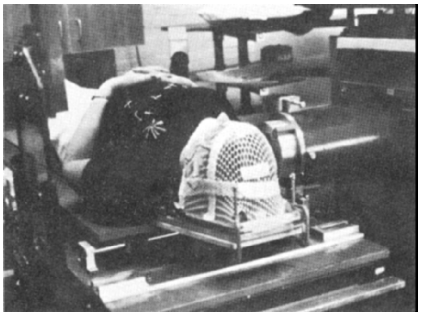


Figure 7.3. The first use of perforated thermoplastic masks was for proton therapy. Since then more elegant commercial versions are now widely available. Reproduced with permission from Verhey *et al.* (1982).

Whole body support

An individualized whole body support can be made from a plaster cast. This is an effective approach, but is somewhat labor-intensive and has largely been replaced by the use of a plastic bag, filled with foam pellets, upon which the patient lies. Once the patient is in the desired treatment position, the bag is made stiff by pulling a near-vacuum on it. There is a danger that the will be punctured, in which case it will lose its vacuum and, hence, shape, and the whole scanning and planning process must be repeated. Loss of vacuum, however, seems to be a rare event. These bags are bulky and require a lot of storage room, but otherwise are very convenient. They automatically satisfy the two-joint rule.

Bite block

The patient's head can be well immobilized through the use of a bite block fixed to the treatment equipment (e.g., to the couch top as illustrated in Figure 7.4. While usually providing good

immobilization, this device suffers from the problem that it is of questionable value in the edentulous patient and may place some strain on the patient. These problems can be largely overcome through the use of vacuum suction on a bite block made to conform to the patient's palate.

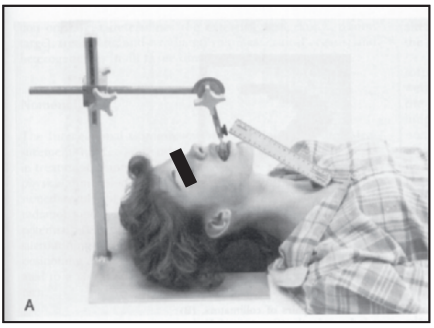


Figure 7.4. Immobilization using a dental bite block.

Stereotactic head holder

A stereotactic frame is a cube-like frame which surrounds and is attached to the head by pins set into burr-holes made in the skull. It serves both to immobilize the head (when it itself is held firm) and to provide fiducial landmarks relative to which the target volume can be located in imaging studies and for treatment. This device was first used in external proton beam radiotherapy for the irradiation of intracranial targets such as the pituitary gland (Kjellberg *et al*., 1962). Similar but noninvasive devices, where frame fixation is accomplished through the use of a bite block and pressure points, are also used now.

When an immobilization device is used, one has to be aware that it may affect the patient's reaction to the delivered dose. For example, some skin-sparing is lost when a mask is used, and the use of a bite block may exacerbate mucosal reactions to radiation.

LOCALIZATION

Once the patient has been adequately immobilized, the next step is to locate the target volume in space, relative to the treatment equipment. This requires that: a) the patient be located reproducibly relative to the treatment equipment; and b) the target volume be in a known spatial relationship to the patient. The latter is generally based on imaging studies, as discussed in Chapter 3. There are four general approaches to localization that are described below.

Localization based on skin marks

In some not very common circumstances such as cancers of the skin, lip, etc., when the target volume is relatively superficial, the target

volume is best localized by localization of the overlying skin. The normal method of localization would then be to adjust the patient's position until a light field coincident with the radiation beam is aligned with tattoos or semi-permanent marks on the skin, the marks having been previously placed there on the basis of a CT study, radiographs taken during simulation, or observation and palpation.

Skin marks are also employed for deeper tumors provided that the patient immobilization is adequate and the tumor can reasonably be expected to remain in a fairly constant relationship with the skin. It is usual to employ laser beams directed along three orthogonal directions. The patient location is adjusted until these beams align with marks placed on the patient's skin at a previous simulation.

Recently, great progress has been made in stereo-photogrammetry and devices are now commercially available that, using optical methods featuring at least two digital cameras, can measure with millimeter accuracy a patient's skin surface in 3D relative to, say, the treatment couch top – and hence, to the treatment machine as a whole. The location in space of the skin surface measured at one time (e.g., just before treatment) can be quantitatively compared with that measured at a previous reference time and the difference between them quantified and, even, corrected for by moving and possibly rotating the patient according to computer-calculated corrections. This method is fast and practical. Figure 7.5 shows an example of such a device and a surface image obtained with it.

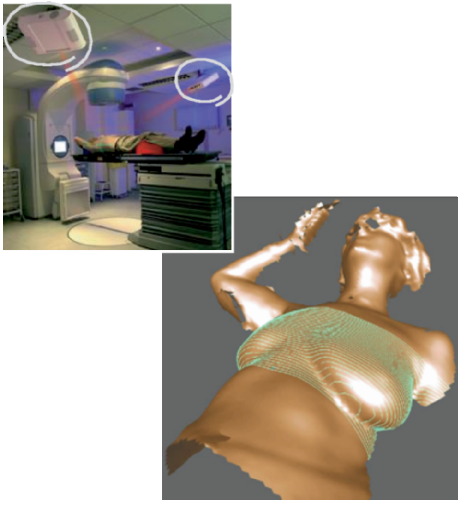


Figure 7.5. Top left: view of a treatment room outfitted with two (for a more complete view) stereo-photogrammetric cameras (outlined), and, bottom right, a projection view of a measured 3D skin surface with computed surface contour lines superimposed. Figure courtesy of Vision RT, Ltd.

Localization based on bony anatomy

It is common in high-precision work to relate the target volume to bony landmarks rather than to skin marks. In this case,

laser beams and skin marks are only used as a preliminary step in the localization process.

The localization of the target volume relative to the treatment equipment based on bony anatomy proceeds in two steps. First, the target volume is located relative to the bony anatomy; and second, the bony anatomy is located relative to the treatment equipment. The first step is accomplished in the treatment planning process, based on the planning CT and other imaging studies. Once the PTV has been delineated, the planning process determines the beams to be used, the central axes of which are generally aimed toward a point within the patient which is located near the geometric center of the target volume. The planning process then establishes the location of the aiming point(s) relative to selected features of the bony anatomy.

The most common way of locating the bony landmarks relative to the patient support system is to compare alignment radiographs taken in the treatment room with digitally reconstructed radiographs computed from the same viewpoints. The localization process generally consists of moving the patient until the anatomy seen in an orthogonal pair of radiographs has the same spatial relationship to the treatment equipment as in the pair of corresponding DRRs prepared as part of the treatment plan. In particular, the location of the bony anatomy relative to a cross-hair that establishes the coordinate system of the radiograph is required to be the same in the alignment radiograph(s) as in the DRR(s) as illustrated in Figure 7.6. The process of establishing this correspondence can be done manually or, more objectively and in principle faster, by using a computer.

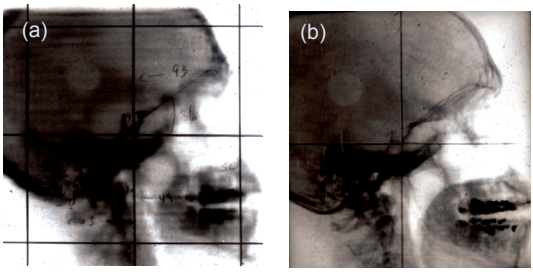


Figure 7.6. (a) A lateral DRR generated by the treatment planning program; and (b) a lateral radiograph taken with the patient in the treatment position. The patient's position has been adjusted so as to bring the anatomy into the same relationship with the cross-hairs in both images.

In a recent development, a set of cone-beam computed tomographic images can be acquired, using either an X-ray tube mounted on the treatment gantry in conjunction with a flat-panel detector or a nearby free-standing cone-beam CT scanner (Jaffrey, 2003). These images

can be compared with the CT image set used for planning, and the geometric differences between the bony (or other) anatomy in the two studies can be used to compute a positioning correction, using the techniques of image registration described in Chapter 3.

Localization relative to the immobilization device

Fiducial markers embedded in the immobilization device can be used in the same fashion as bony landmarks. Because fiducial markers can generally be located very accurately, localization in such cases can be more accurate than when using bony landmarks – provided that the patient is securely and reproducibly held in the immobilization device as, for example, when a stereotactic head holder is used. It is important that this proviso be satisfied; otherwise, there is a danger of treating the immobilization device rather than the patient.

Localization based on identification of target markers or the tumor itself

In some circumstances, radiographically visible objects such as gold seeds or surgical clips can be deliberately or fortuitously embedded in or close to the tumor. These can be radiographically localized – and even tracked during radiation delivery. For example, gold seeds have been introduced into the prostate for the purpose of tumor localization (Shipley *et al*., 1979) and tantalum clips are sutured to the sclera of the eye in treating uveal melanoma as described in Chapter 11. In some cases, the GTV itself may be visible, for example when using ultrasound to locate the prostate just before treatments. Such techniques provide accurate target volume localization. The localization process in the case of radioopaque markers follows that for bony landmarks.

VERIFICATION

Once the patient has been positioned for treatment, it is desirable to verify the alignment of the beam relative to the target volume. It may also be desirable to determine, after the treatment has taken place, whether or by how much the patient has moved during treatment. Comparison of "before" and "after" measurements can provide valuable information on the efficacy of the immobilization techniques (Verhey *et al*., 1982; Verhey and Bentel, 1999).

Verification using portal radiographs

The most direct method of verification for photon treatments is to use the therapy beam itself to make a radiograph of the anatomy through

which it passes. In a short exposure, the transmitted beam will make an imprint on a film or digital imaging plate of the anatomy included within the beam. This radiograph can be compared with a DRR designed by the treatment planning system in the beam's-eye view, and any needed adjustments made before treatment proceeds. Sometimes a double exposure is made consisting of one exposure of the treatment field and a second with the field opened up – for example by opening up the multi-leaf collimator jaws. A double-exposure film provides visualization of anatomic landmarks close to but outside the beam, but has the disadvantage that radiation is delivered to tissues outside the target volume which do not require it. Unfortunately, the quality of radiographs made by photons of therapeutic energy is very inferior to those made with diagnostic X-rays, as alluded to in Chapter 4.

Verification using X-radiography

It is feasible to employ a pair of X-ray tubes mounted in the treatment room, in a known relationship to the treatment equipment, and directed toward the isocenter. These X-ray tubes need not necessarily be directed orthogonally to one another (Schweikard *et al.*, 2004). The radiographs thus obtained can be compared with DRRs from the same radiographic viewpoints, computed by the treatment planning program. It is also possible to provide fluoroscopic imaging and hence real-time localization during treatment to adjust the position of the patient relative to the beam in real-time.

ORGAN MOTION

Organs and tissues both move within the body and change size and shape, both during the delivery of a single fraction (intra-fraction motion), and over the course of the entire therapy (inter-fraction motion). Motion poses a number of problems, chief among which are: (1) the imaging study or studies upon which the treatment plan is based (that can either be at a single moment in time, or an average/distortion over time) are inaccurate and hence give a false picture of the anatomy; (2) larger fields are needed than the size of the CTV would seem to require, and hence more normal tissue is irradiated than would otherwise be necessary; and (3) if the extent of the motion is not fully appreciated, fields may be designed too small with the danger that parts of the tumor my be underdosed.

Inter-fraction movement, including size and shape changes, of the tumor and/or organs may take place on a day-to-day or week-to-week

basis, being caused, for example, by changes in bowel or bladder filling, tumor regression, changes in the patient's weight and so forth.

Intra-fraction motion may occur on a range of time scales. Motion caused by the beating of the heart is quasi-periodic in nature, with a cycle time of about 1 second; motion caused by respiration is quasi-periodic, with a cycle time of about 4 seconds; motion caused by peristalsis is aperiodic and can take place over time scales of up to a minute. Of these motions, respiration is probably of greatest importance. Respiration can sometimes result in excursions of organs of a few centimeters, even when the organ is some distance away from the diaphragm (e.g., the kidney).

The impact of organ motion on imaging

The problems caused by organ motion arise during imaging, planning, simulation, and treatment. During simulation using X-radiographs, the images, while sharp, may not be representative of the tumor position, since they are a single short exposure taken at one moment in time during the breathing cycle.

For many other forms of imaging, including CT simulation, the reconstruction of the scans may be distorted due to image blurring and synchronous effects between the rates of image sampling and respiration (Chen *et al*., 2004). CT techniques have been developed, using single-slice or multi-slice CT scanners and respiration monitoring, that have made it possible to obtain multiple sets of CT images that are correlated with fairly well-defined phases of the respiratory cycle. These so-called 4DCT scans can be used to select phases of the breathing cycle where motion is at a minimum and, by turning off the radiation beam during those phases, smaller safety margins can be employed and, hence, less normal tissue irradiated.

Organ motion in the absence of special measures

Motion of the patient as a whole will, of course, result in motion of internal structures. This motion is minimized by adequate immobilization of the patient, as discussed above, and will not be futher discussed here.

So far as organ motion within the patient is concerned, Langen and Jones (2001) have reviewed a number of studies which have documented the extent of motion of several organs. Typically, the extent of motion can vary from a negligible amount, to excursions of a few centimeters or so in tissues near to, or influenced by,

diaphragmatic movement. In the absence of special measures, the only way to deal with situations in which large excursions can occur is to allow generous margins in delineating both the PTV(s) and PRVs. It is quite possible that, where the extent of motion or of the artifacts that it produces has been underappreciated, the probability of local control has been compromised (Ling *et al*., 2004).

Organ motion with respiration gating

The most obvious and simplest way to handle respiratory motion is to track the respiratory cycle, identify the phase(s), usually during expiration or quiet breathing, where motion is least, and turn the beam off (gate the treatment) during the other phases.

Respiratory gating (Ohara *et al*., 1989) uses an external breathing monitor to gate the radiation beam on and off at a well-defined phase of the breathing cycle. An example of such a monitor is a light emitting diode or other optical target, placed on the patient's abdomen, whose position is monitored by video cameras, while the patient breathes. The diode position may be used to gate CT or fluoroscopic data acquisition, and may be used during treatment to gate the accelerator beam, thus reducing the effect of respiratory motion by synchronizing the dose delivery to the patient's breathing cycle. A wide variety of position monitoring devices have been used, including a strain gauge or linear transducer attached to the abdomen or thorax, or a temperature sensitive device inserted in the nostril. Stereo-photogrammetric cameras, mentioned above, can also be used to monitor respiration.

There are also approaches that seek to actively control the flow of air to the patient, and hence tumor motion. These include: deep breath hold at inspiration controlled by the patient, viewing a signal from a spirometer; and active breathing control, an approach in which the patient breathes through a mouthpiece connected to a pair of flow monitors and valves which are closed at a preselected phase in the respiratory cycle.

Organ motion with tumor tracking

A problem with the above techniques is that they reduce efficiency, as radiation can only be delivered during a portion of the patient's breathing cycle, or the irradiation must be interrupted between breath holds. In addition, they rely on measurements of the relative positions of the tumor and the normal anatomy which are made well in advance of treatment. These measurements are assumed to apply at

the time of treatment, but there is evidence that this may not always be the case. A more elegant solution would be to track target motion during the treatment (i.e., by imaging implanted seeds or surgical clips) and adjust the position of the beam relative to the patient appropriately, while the accelerator runs continuously. The adjustment could be achieved by moving the patient couch or by moving the radiation beam – for example, by adjusting the settings of a multi-leaf collimator to track the tumor as it moves (that has the added advantage that it could compensate for shape changes, if they are known).

Correlation of tumor position with phase of respiration

The various methods of gating or breath control all have the advantage that the extent of motion of tumors and organs due to respiration can be substantially reduced. The only fly in the ointment is the extent to which tumor position correlates with the respiration monitor or method being used. This is a matter of intense investigation at the time of writing. Observations have been made that call the exactness of the correlation into question. However, I believe that, in the majority of instances, breath gating can significantly reduce the amount of motion and permit tighter field margins to be used, even if they are not as tight as might be possible with more complex approaches such as tumor tracking.

COMPENSATION FOR PATIENT AND ORGAN MOTION

For any given set of patient immobilization and patient and organ localization techniques, there always remains some degree of residual motion and some uncertainties about where the patient, target volume(s), and OARs are located relative to the treatment couch top. These uncertainties must be taken into account in planning the treatment.

Adding lateral margins to the beam

When a uniform beam is directed at a target volume with the intention of irradiating the entire target volume, there are two factors that affect the size that the beam must have, namely:

1. Motion, in a broad sense, of the CTV with respect to the beam that has two components: intra-fractional patient and organ motion that gives rise to the ITV; and equipment set-up errors that, added to the ITV, constitute the PTV (see Chapter 3).

2. The inherent penumbra of the beam that requires that one must make the size of the beam, defined as the geometric shadow of the collimators or aperture, larger than the projected size of the PTV.

In effect, a margin must be added all around the field to allow for these two factors. The margin needs by no means be uniform; rather, it should reflect the possible degree of uncertainty at each point of the field periphery. A common approach is to enlarge each beam in a plan so as to place the 95% isodose of the combined beams at the PTV edge.

A further form of "motion" consists of anatomic changes that may occur over the course of the treatment (so-called inter-fractional changes). Allowance for the possibility of such changes needs to be built into the field margins, although the size of the added margin may be reduced by periodically re-evaluating the patient's anatomy and tumor.

The influence of neighboring normal tissues

As already alluded to, if normal tissue damage were not an issue, one could make the beam simply enormous and thus ensure CTV coverage. But, of course, the irradiation of normal tissues *is* a critical issue. *The field margins must be chosen so as to achieve a balance between local tumor control and morbidity – that is, between TCP and NTCP.* In principle, then, this requires an understanding of the biology of the tumor and of each normal tissue of importance. To help understand how this balance may be made, I present here the results of a simple modeling exercise.

The basis for choosing safety margins – a simple model

The calculation relates to a clinical target volume irradiated by two parallel-opposed beams and analyzes the dose profile across the tumor in the face of tumor motion. For simplicity, the model is two-dimensional. The beam penumbra on each side of each field is assumed to have the shape of an error function with standard deviation, p. The 50-50% width of the dose distribution, w, is chosen so that, in the absence of motion, the target volume periphery receives 95% of the central axis dose. The beam, however, is assumed to move relative to the target volume with a Gaussian distribution of motion having a standard deviation of m. To compensate for this motion, the beam is widened by a safety margin, s, these parameters are illustrated graphically in Figure 7.7a.

The question is, what should the margin, s, be? In a sophisticated calculation, the normal tissue tolerance of each nearby normal tissue would be taken into account separately. In the model whose results are presented here, all normal tissues are lumped together and the assumption is made that, as the irradiated volume increases, the central axis dose must be decreased to keep morbidity at the same level, using the volume dependence discussed in Chapter 5, namely

$$\text{central axis dose} = \text{prescription dose} \cdot [(w+2s)/w]^f$$

where f is the volume dependence factor, here taken to be –0.1 and the factor $(w+2s)$ is the beam width enlarged to allow for motion. With these assumptions, for a 10 cm diameter tumor, one has the result shown in Figure 7.7b in which the estimated EUD is plotted as a function of the safety margin, s, given in units of $\sqrt{(p^2 + m^2)}$.

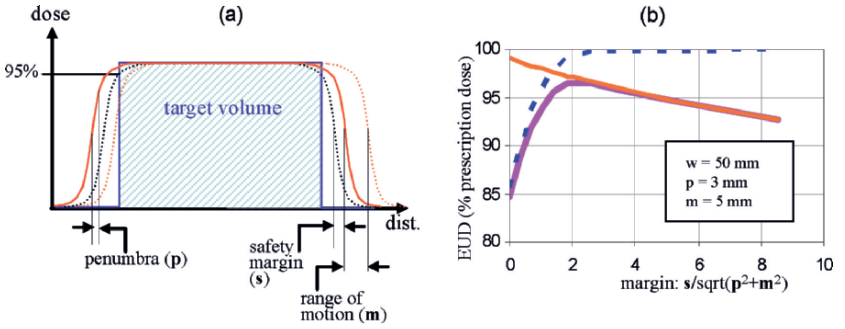


Figure 7.7. Model of EUD as a function of beam margin. (a) schematic illustration of the parameters involved. The *dotted black curve* is the beam profile in the absence of motion. The *solid red curve* is the beam profile after having added a safety margin, s. The *dotted red curve* is an example of an offset beam profile due to motion. One sees the reduced dose which it gives rise to on the left-hand side of the figure. (b) A plot of EUD (based on an EUD parameter of –10) vs. the added safety margin, s, in units of the sum in quadrature of p and m (see text).

In Figure 7.7b:

The dotted blue line shows what happens if one increases the beam width (by increasing the safety margin, s) without lowering the dose to keep the normal tissue reactions the same. The EUD drops when too small a margin for motion is allowed, since the target edge will then be underdosed, and rises to 100% of the prescription dose as the field is enlarged, thus ensuring that the prescription dose is given to the entire tumor.

The downward sloping orange line shows what happens in the absence of motion. One sees only the consequences of reducing the dose as the field is widened, in order to maintain the same level of toxicity. The EUD progressively reduces as the field is made wider and the dose is reduced.

The hump-shaped purple line shows what happens when both effects are allowed for. It is, in essence, the product of the two other curves. For too-small margins, the EUD is low due to underdosage at the target edge. The EUD then rises as the margin increases and the target coverage is thereby improved. Finally, at too-large margins, the EUD falls due to the need to reduce the dose in order to keep the normal tissue morbidity constant.

Conclusions from the model

This model is very simplistic and its results should not be taken as being quantitatively accurate. However, it illustrates a very important point, namely that *there is an optimum margin size, that gives the highest EUD (and hence TCP) for a given fixed level of normal tissue toxicity*. Smaller or larger margins would be worse, i.e., would lead to lower EUDs. The model also predicts that *the optimum margin is approximately two times the standard deviation of random motion* (slightly corrected for the penumbra size). Figure 7.7b captures the essence of the basis for choosing the best safety margin to use.

Random and systematic motion

Motion may be either random or systematic. If random, variations in position occur during a patient's treatment – either between fractions, or more usually, within a treatment fraction – with a Gaussian-like distribution of values. Systematic motion, on the other hand, is likely to show up as a consistent error in a patient's treatment, that may or may not vary from patient to patient. The distinction between these two types of uncertainty in the location of patient anatomy was introduced in Rabinowitz *et al*. (1985) where simulator and port films were analyzed. Retrospective analysis of day-to-day variations in the location of anatomic landmarks or metal clips relative to the field border were seen to be approximately Gaussian in distribution and were interpreted as random variations. On the other hand, there were consistent deviations between the port films and the initial simulation film (that was taken to represent the desired beam placement) and these were interpreted as systematic variations. Of interest was the

fact that the systematic variations were, at most anatomic sites, larger than the random variations. These observations have been repeated in many subsequent studies.

Figure 7.8 illustrates, in a simple example, the way in which random and systematic variations differ so far as treatment outcome is concerned. In this figure, a tumor is being irradiated by parallel-opposed fields just large enough to cover the tumor if the fields were correctly aligned. However, the fields are not correctly aligned, being shifted half the time in one direction by, say, a distance of 20% of the tumor diameter, and the other half of the time by the same distance in the opposite direction.

In the case of random motion, all patients receive the dose illustrated in Figure 7.8a, namely the prescription dose in the central 80% of the tumor, and 50% of the prescription dose at the two sides. While this dose is by no means ideal, it carries a finite chance of tumor control. On the other hand, in the case of systematic motion, all patients will have zero dose over 20% of their volume, as seen in Figure 7.8b, and none will be controlled.

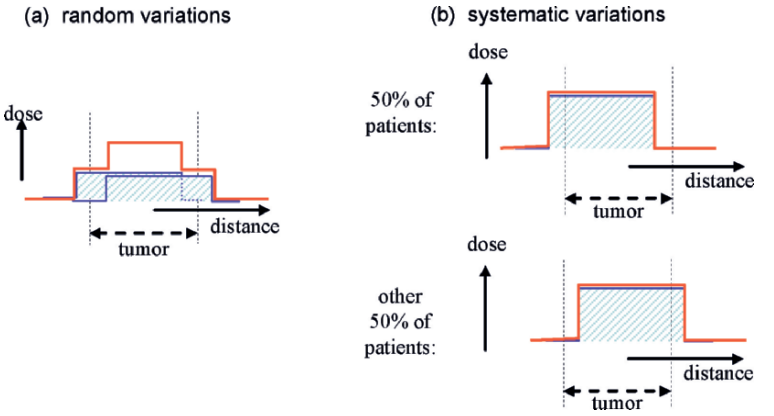


Figure 7.8 Schematic comparison of two types of motion.
(a) random motion, where half of the time during an irradiation the beam is to the right and half the time it is to the left. The composite dose distribution is shown in *red* (the dose profiles are slightly vertically staggered for clarity).
(b) systematic motion, where half of the patients receive no dose on the left side and the other half receive no dose on the right side.

While this example is highly simplistic, it illustrates the general point that random positioning errors smear out a given patient's dose distribution, whereas systematic positioning errors can leave each

patient with a more significant dose deficit. As a result, systematic positioning errors have a greater impact on TCP and hence require larger margins, than random errors of the same magnitude.

Detailed models of the required safety margin

Much more sophisticated ways of selecting margins to compensate for motion have been developed and a good review of many of these can be found in van Herk (2004). One example of a margin prescription is the following (van Herk *et al*., 2000):

1. estimate the total random uncertainty, σ, by adding the estimates of random uncertainty from all sources together in quadrature;
2. estimate the total systematic uncertainty, Σ, by adding the estimates of systematic uncertainty from all sources together in quadrature;
3. estimate the necessary safety margin that needs to be added as $2.5\Sigma + 0.7\ \sigma$.

This recipe was designed to ensure that 90% of patients have a minimum dose to the CTV of 95% of the prescription dose. A weakness of this approach is that the margin prescription is based purely on target volume considerations, without reference to normal tissue complications.

SUMMARY

Motion and mis-registration of the target volume with respect to the radiation beams is, at some level, inevitable. If the target volumes are to be adequately irradiated, and adjacent OARs are to be protected, it is essential that:

1. the causes and possible magnitudes of motion and misregistration (in ones own institution) are understood;
2. their possible consequences are understood;
3. measures be taken to minimize motion and mis-registration to the extent possible and clinically warranted; and that
4. steps are taken (primarily through the provision of judiciously chosen field margins) to allow for the remaining degrees of and mis-registration.

8. PLANNING MANUALLY

INTRODUCTION

Finally we have reached the point at which we can discuss the actual planning of a treatment. The preliminaries have been taken care of: we have the imaging studies and the needed volumes of interest delineated; we know how a single photon beam is constructed; and we have the planning aims in front of us. Then, too, we know how to evaluate any plan we devise, and how to compare it with alternative plans. Let the fun begin!

There are two basic approaches to developing a plan. The first is what I call *manual planning*. This is by no means all manual, since computers and graphical displays are heavily used. What is "manual" is the way in which the plan is assessed as it is iteratively improved. The assessment uses what I called "expert inspection" in Chapter 6, and is a largely subjective process which is based on a review of a very large number of computed quantities. Manual planning usually, but not always, has as its goal the development of uniform-intensity radiation therapy.

The second approach is what I call *computer-driven planning*. By this is meant that decisions about a plan's worth are made by computer. This is done out of necessity, because in computer-driven planning a huge number of plans are tried out. Computer-driven planning usually, but not necessarily, has as its goal the development of intensity-modulated radiation therapy and virtually always uses the optimization techniques discussed in Chapter 9.

PLANNING BY HAND

Figure 8.1 depicts a planner sitting in front of a console that has a large number of knobs and wanting to decide on the setting of each knob in order to arrive at an acceptable treatment plan. How is this poor guy to adjust all the knobs so as to arrive at even a good, let alone optimal, plan?

Figure 8.1. The problem of planning therapy: so many "knobs" (treatment variables) to tweak! How can the planner decide how to set them so as to deliver a passable, let alone optimal, treatment?

The knobs, of course, are the variables that can be adjusted. The following is a partial list of those variables:

- the type of therapy (external beam, intracavitary or interstitial implant, intraoperative – or, some combination of these). If external beam therapy, then:
- the modality of any external radiation beams (e.g., X-rays, electrons, protons etc.) and the characteristics (e.g., the energy) of the chosen modality;
- the location of the patient and the tumor and organs of interest within the patient in both space and time – including measures to control, or to make allowance for, uncertainties in organ and patient location relative to the treatment beams;
- the number of external radiation beams;
- the angulation and aiming point of each beam;
- the shape of each beam;
- the weight and intensity profile of each beam;

The choice of all of these variables, most of which are dictated by purely clinical considerations, is at the heart of the planning process.

Manual planning is the approach that has been taken since the very beginnings of radiation therapy more than a hundred years ago. It takes advantage of:

- the memory of previous satisfactory knob settings (that is, of plans used previously for similar cases) as a starting point;
- rules of thumb as to how to set combinations of knobs – e.g., "design the aperture so as to have the beam just cover the target volume[1] with a predetermined margin or margins;"
- a fast calculation engine to compute, ideally interactively, the dose distribution resulting from a particular set of knobs;
- displays of that dose distribution;
- the provision for the planner's inspection of a number of dose-summarization statistics – e.g., dose–volume histograms and/or calculation of the minimum, maximum, and mean tumor dose, and so forth;
- a body of experience that makes a judgment about the overall acceptability of the plan;
- the iteration (a few times) of the process to arrive at the best plan the planner is able to come up with.

A glimpse of a planner engaged in manual planning is shown in Figure 8.2. You will surely observe that this drawing is, on the face of it, a rather simple extension of Figure 8.1. Apparently, all we have to do is to connect the knobs to a calculation engine and show the resulting dose distribution, and other quantities derived from it, on a screen. What you do not see, because his

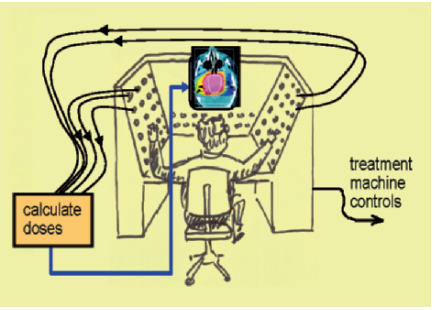


Figure 8.2. Planning treatment using manual planning (see text).

[1] To avoid lengthy qualifications, I have used the generic term "target volume" throughout this chapter, without specifying whether the GTV, CTV, or PTV (see Chapter 3) is meant. Generally, however, the PTV is implied.

back is turned to us, is the look of intense concentration, possibly even desperation, on the planner's face. This is because the central requirement of the manual refinement process is that the planner has to "make a judgment about the overall acceptability of the plan." This is the difficult problem of plan assessment that has already been discussed in Chapter 6.

The flow chart that describes what our planner is up to is shown in Figure 8.3. It is a simple iterative loop in which the planner starts with some approach (i.e., a group of knob settings) and evaluates the plan that would result. Then, based on experience, he adjusts the knob settings and tries again, and again… until he is satisfied.

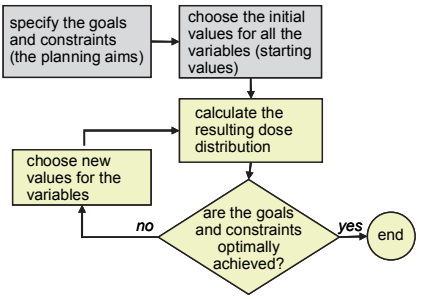


Figure 8.3. Flow chart for manual planning (see text).

Developing a manual plan

Our planner is not entirely without some arrows in his quiver. These include the following.

Choice of radiation modality and beam energy

There can be several modalities from which to choose. As regards external beam radiation therapy, these include photons, electrons, protons, and so forth. They, of course, have entirely different characteristics and the planner, when more than one modality is available, can take advantage of these differences. Table 8.1 summarizes some of the more important clinical differences between single beams of different modalities.

In addition, for each of these modalities, one can chose (or, in the case of the charged particle beams, calculate) the desirable beam energy.

Design of field shape

Once a provisional direction for a particular beam has been chosen, one needs to design the shape of the field. That is, one must decide on the collimator settings, and design the aperture and/or blocks or multi-leaf collimator settings that will block out parts of the otherwise

Table 8.1. Comparison of principal advantages and disadvantages of beams of various modalities used in external beam radiation therapy.

	advantages	*disadvantages*
photons	widely available good skin sparing	higher entrance dose than tumor dose high dose throughout patient up to exit surface
electrons	finite penetration, thus sparing tissues distal to the target volume very slight skin-sparing	broad penumbra due to scattering only suitable for quite shallow target volumes due to shallow fall-off of the distal dose at higher energies
protons	virtually no dose distal to the target volume somewhat reduced entrance dose proximal to the target volume	management of inhomogeneities is non-trivial Penumbra becomes substantial at large depths (e.g., ≥ 20 cm) no skin sparing very limited availability

rectangular field. Here, the goals are largely clear. They are: (1) to cover the entire CTV; (2) to do so with adequate margins to take care of the various factors relating to patient and organ motion and setup errors, and the beam penumbra as discussed in Chapter 7; and (3) to avoid any OARs of particular concern – or, if one is not entirely avoidable, then to minimize the volume of the OAR that is covered by the beam.

A particularly useful approach is the design of the field shape in the beam's-eye view (BEV). The beam's-eye view is a perspective view of the patient's delineated anatomy as seen from the viewpoint of the radiation source of one particular beam. As the planner changes the beam direction the display changes, showing the new spatial relationships between the target volume and the delineated anatomy This allows the planner to choose a beam direction from which particular OARs can either be avoided, or minimally included in the

beam. An example is given in Figure 8.4, where the beam direction has been decided upon and the beam aperture is being drawn, making use of a circular cursor that helps the planner to leave a specified boundary around the CTV.

Figure 8.4 is an old and even historic image (Goitein *et al*., 1983). Modern graphics engines give much more attractive surface-rendered images, but they do not really make the process any more accurate or easy. Note, also, the equipment settings, displayed in white below the BEV, and the three orthogonal views of, in this case, the proton treatment apparatus which includes a treatment couch with six degrees of freedom. In this system, as in later planning systems, the planner could adjust the beam settings using those variables that are intrinsic to the equipment, such as couch height, gantry angle and so forth, providing what has since been termed "virtual simulation."

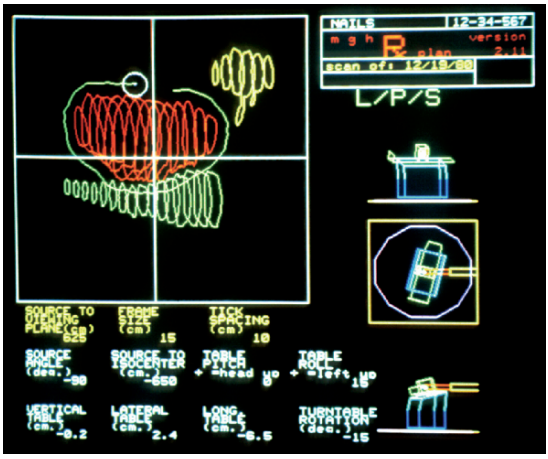


Figure 8.4. Beam's-eye view of patient anatomy. A beam aperture is being drawn with the aid of a cursor whose radius equals the desired margin. Reproduced with permission from Goitein *et al.* (1983).

Computer tools exist to design the aperture automatically, given the desired field margins (that need not be the same all around the PTV).

Decision as to how many beams to use

Generally, the choice of the number of beams depends critically on the patient's individual geometry. One central decision, further discussed below, is whether to use arc therapy, which involves rotating the beam around the patient over a range of angles up to the full 360° possible, or to employ a few fixed beams. Rarely, as has already been discussed, would a single beam be a good choice except in the case of superficial tumors. Generally, too, a pair of parallel-opposed photon beams gives too high a dose outside the target volume to be useful. Typically, then, a few beams – say between 3 and 7 – are chosen. Plans with fewer beams are sometimes preferred

because of the greater ease and, perhaps, safety with which they can be delivered.

Determination of beam direction(s)

Generally, it makes little sense to use two photon beams that cover the same volume but are separated by only a small angle, say 15° or less; their combined dose distribution will not be very different from that of a single beam.

Quite frequently, the geometry of the target volume and/or of a critical OAR will suggest a particular approach. This is the case, for example, with the beam illustrated in Figure 6.9 in Chapter 6 where the obliquity of beam direction has been chosen so that the beam's edge runs near-parallel to the medial surface of the target volume.

Naively, one might think that all beams should avoid all OARs.[2] However, this is never possible; one cannot achieve zero or very low dose except in a limited volume. One then makes a conscious decision to include a particular OAR in one or more beams, subject only to the requirement that the dose thus delivered be below the constraints established in the planning aims. The directions of other beams are then picked to avoid the OAR in question to the extent possible.

Modern linacs have a rotating gantry that allows beams to be directed towards a central point (isocenter) from any direction within a plane perpendicular to the axis of the gantry's rotation (see Figure 1.1 of Chapter 1). The patient couch, in addition to being able to move the patient laterally, longitudinally and vertically, can rotate in a horizontal plane about isocenter, thereby enabling the use of beams directed out of the transverse plane. However, this degree of freedom is rarely used[3] and most plans used in practice feature coplanar beams

---

[2] The concept that there are radiosensitive and non-radiosensitive OARs, and that only the sensitive ones need to be worried about, is somewhat dangerous. All tissues are affected by radiation and one should not ignore the dose to any OAR.

[3] In earlier times, when I visited another radiotherapy department, I would often ask that the treatment couch be rotated off the straight-ahead position. Almost inevitably there was a pile of dust where the couch had been, suggesting that it had not been rotated for a long time; otherwise the cleaning crew would have seen the dust and taken care of it. The

[continued on next page

that lie in a plane. This is a great pity because a noncoplanar approach often has advantages. In our experience with proton beam therapy at the Harvard Cyclotron Laboratory, we found that we employed non-coplanar approaches in about two-thirds of the treatments.

Determination of beam weight(s)

Not all beams need to be equally weighted – that is, need to deliver approximately the same dose to the target volume. Often, one chooses to weight some beams more heavily than others. Just how the weighting is decided upon is a matter of experience and trial and error, involving expert judgment – bolstered sometimes by rules of thumb developed in the clinic. The optimization techniques described in Chapter 9 for IMRT can equally be applied to the problem of choosing beam weights in uniform-intensity radiation therapy (Niemierko, 1992).

Iterations of the planning process

Rarely does a satisfactory plan, even when developed by a very experienced planner, emerge after the first attempt. The exception to this is when standard planning approaches (class-solutions) are required by protocol. Normally, the planner will work on several plans, choose the two or three best, and consult with the clinician as to which he or she prefers.

ENVIRONMENTALLY FRIENDLY DOSE DISPOSAL

Let me start at the outset of this section to state what I believe to be the central tenet of treatment planning. Namely, that *the planner's job is to decide how to dispose of (i.e., distribute) the dose that must inevitably be delivered outside the target volume in the best manner possible*. The terms "dose dumping" and "dose littering" have also been coined for this phenomenon. Planners are, in the last analysis, disposal engineers.

---

immediate conclusion was that noncoplanar beams were probably used at best infrequently at that facility.

Integral dose

Dose deposited outside the target volume is the toxic substance that has to be disposed of. Integral dose is a measure of how much toxic material is involved.

Integral dose is estimated by dividing the tissues outside the target volume into small subvolumes, in each of which the dose is approximately uniform, multiplying the dose in each subvolume by its mass,[4] and then adding this product up for all the subvolumes. Since the units of dose are energy per unit mass, one can easily appreciate that the product of dose in a sub-volume and its mass measures the energy deposited in that sub-volume. *Integral dose is then the measure of the total energy deposited in the patient outside the target volume*.

Integral dose, *per se*, does not directly correspond to tissue damage. However, it is a very useful quantity to use for accounting purposes. The planner's task as just defined is to distribute this energy with the minimum deleterious consequences for the patient.

Impact of treatment approaches on integral dose

It is interesting to see how the type of treatment affects the integral dose. Presumably, treatments that generate less integral dose and are otherwise acceptable are likely to be better for the patient. Figure 8.5 is a highly schematic representation of a number of treatment approaches. For each panel, a very crude estimate of integral dose is given, assuming that the same dose distribution applies in other parallel sections of the patient.

As Figure 8.5 suggests, in many cases the treatment technique has very little influence on the integral dose delivered. Neither the number of beams (panel a) nor their relative weightings (panel b) has much impact on the integral dose, nor does the energy of photon beams, except for very large patient cross-sections. Although not shown here, the observation that beam weighting does not much influence the integral dose carries over into IMRT where it is the internal weightings between pencil beams, rather than the relative weightings of the whole beams, that have little effect on integral dose.

[4] In practice, it is usually the volume and not the mass of each subvolume which is used – on the grounds that most soft tissues have near unit density.

In addition, modest differences in integral dose occur when the body outline is non-spherical, in which case beams passing through thick portions of the body deposit dose in a larger volume of tissue. Similarly, modest differences in integral dose occur when the target volume is non-spherical in which case beams directed along the direction in which the projection of the target volume is small deposit less dose because the beams need to have smaller cross-sections, as illustrated in panel c.

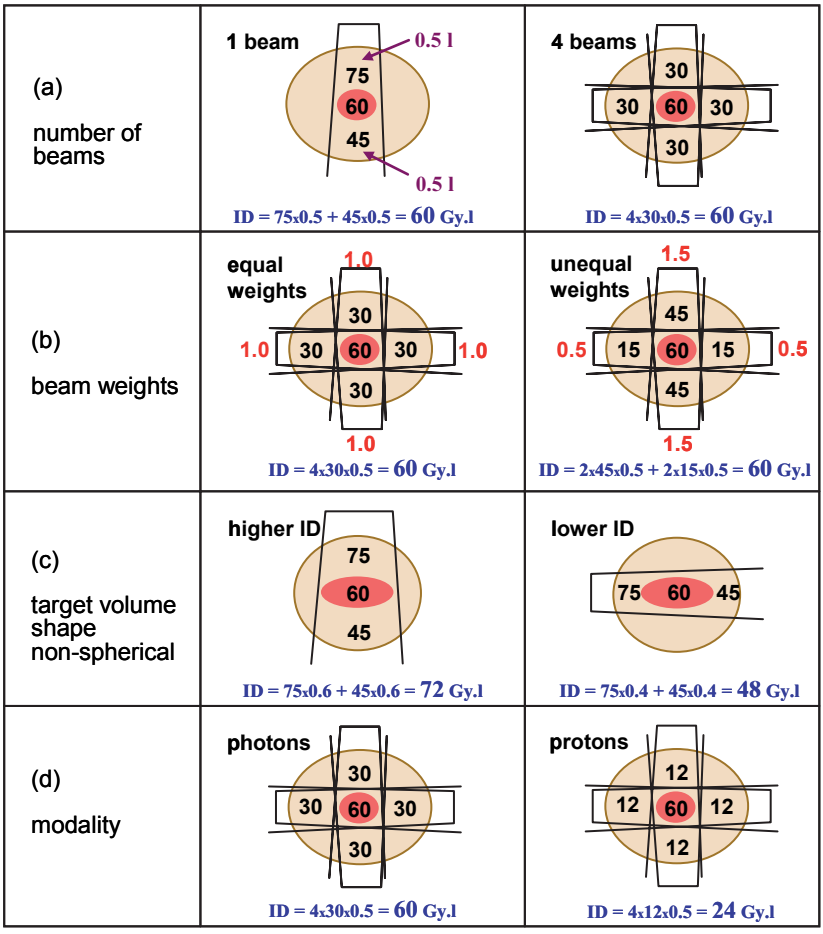


Figure 8.5. Schematic presentation of a number of different treatment approaches and their impact on integral dose (see text). A very crude estimate of the integral dose (ID) is shown under each sketch.

Much bigger changes in integral dose occur when:

- the projection of the target volume is irregular, so that a rectangular field can be reduced in area by employing a custom-made aperture to block the unnecessary parts of the periphery of the field;
- the target is close to the patient surface, so that fields can be used used that enter the patient close to the target; and
- when a different radiation modality is employed – specifically charged particles such as electrons or protons. Panel (d) of Figure 8.5 shows this effect, and reinforces the point that the gain in integral dose from using, say, protons, is fairly independent of the number of beams used, just as for photon beams.

Where to dispose of the dose?

Given that a planner has a choice of where to dump the dose, where should it be put? The first, and almost trivial, part of the answer to this question is that the dose should be distributed so that, as far as possible, the treatment aims are achieved. However, the treatment aims are usually expressed for specific organs and therefore, as I have already suggested, are likely to constrain the dose in only part of the total volume irradiated. So the question remains: once the treatment aims have been satisfied, where within the remaining patient volume should one put the remaining beam energy? Of course, one does not have a free choice; the laws of physics set limits in how it can be distributed. In making this decision, the following question is crucial.

A lot to a little or a little to a lot?

One generally has the choice of delivering a high dose to a modest volume of normal tissue, or a lower dose to a larger volume. The integral dose, as we have just seen, is approximately the same in both cases. I like to pose the question as "Is it better to use what is called a 4-field box technique, or a 360° rotation?" The choice is illustrated in Figure 8.6 where, as is usual in dose calculations, the rotation is approximated by a large number of fixed fields.

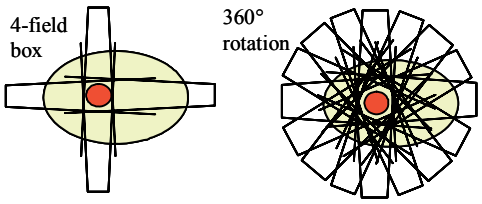


Figure 8.6. The classical choice: should one prefer to use a 4-field box (a lot to a little) or 360° rotation (a little to a lot)?

If one were to plot the dose distributions of the non-target tissues for these two cases as overlaid DVHs, one would get something like the histograms seen in Figure 8.7. This figure raises the vexing problem of crossing DVHs. If a normal tissue DVH for one plan is everywhere below that for another, it is easy to deduce that it is the better (i.e., will be less morbid for the patient) and, equally, the DVH which lies everywhere above another is the worse. But, what is one to think if they cross? In Figure 8.7 the red DVH (for the 360° rotation plan) shows a large volume receiving low doses, the blue DVH (for the 4-field box) shows a lesser volume receiving low doses, but a greater volume receiving higher doses. Which is one to prefer?

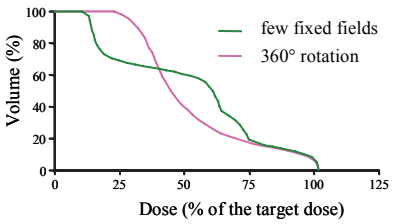


Figure 8.7. Overlaid DVHs for the non-target tissues for two plans.

If I knew a definitive answer to this question I would happily reveal it. But, the fact is that we have here reached the boundary of our knowledge – or, at least, our ability to quantitatively evaluate plans; the clinician's experience is all we have to go on. And, if we cannot answer this question, then what use are all our models, and what faith can we put in a computed score?

Having taken this pessimistic position, let me step back a little and present a modest bit of modeling.

The influence of tissue architecture

Andrzej Niemierko and I some time ago undertook a simple computer experiment. We created a cylindrical tumor of 8 cm diameter within a cylindrical patient of 20 cm diameter, much as for the cases sketched in Figure 8.5, and planned its irradiation with photon beams, in one case with a 3-beam technique and in the other case with a 360° rotation.

The tumor was required to receive essentially the same dose in both cases. The normal tissue outside the tumor received, of course, a quite different dose distribution in the two cases. In fact, the DVHs in Figure 8.7 are the normal tissue DVHs for this experiment. Two types of normal tissue were considered: serial and parallel. We then applied two NTCP models (Niemierko and Goitein, 1991, 1993a) to compute the difference in the NTCP in the two plans, separately for

each type of tissue. The models had several variables and we varied them all. Figure 8.8 shows graphs of the difference in NTCP between the two plans, for the two tissue architectures, and for two of the more significant variables in the analysis in each case.

What one sees is that, in the case of a serial tissue architecture, the rotation plan appears to be better for all combinations of the two variables shown. However, in the case of a parallel architecture tissue, while there are combinations of the two variables for which the rotation plan would be preferred, there are also regions where the 3-field plan is better. These regions are those in which the D_{50} of the tissue is small compared with the tumor dose – i.e., where the parallel architecture tissues are quite radiosensitive – and/or the critical volume is large.

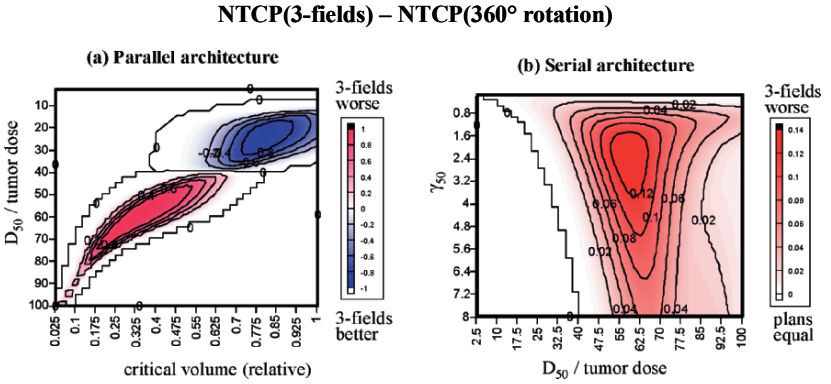


Figure 8.8. Graphs of the difference in NTCP between a 3-field and rotation plan: (a) normal tissues have a parallel architecture, and (b) normal tissues have a serial architecture. Shades of blue are negative differences, implying that the 4-field plan results in a lower NTCP. Shades of red are positive differences, implying that the rotation plan results in a lower NTCP.

One can perform the same kind of analysis using the EUD model described in Chapter 5 which has only one parameter, **a**. In this case, one finds that the rotation plan is preferred for tissues that have $\mathbf{a} > 1$ and the 3-field plan is preferred for the lesser number of tissues that have $\mathbf{a} < 1$. Interestingly, the two approaches are predicted to be equally good in the case of $\mathbf{a} = 1$. For that value, the average normal tissue dose determines the NTCP – and the average normal tissue dose and the integral dose are the same in the two plans.

Of course, this exercise is highly simplistic. For one thing, we are lumping together all tissues outside the tumor and assuming they have the same structure and radiation sensitivity. For another, the models themselves are largely unproven. So, one can certainly not draw any quantitative conclusions from our exercise. However, we did come away with one very important qualitative conclusion, namely *that there is probably no universal answer to the question of which of two crossing DVHs is better – and, therefore, to the question of whether a fixed-field or rotation plan is better – the answer depends on the normal tissue architecture*. Biology matters.

My own inclination is toward the use of a fewer number of beams covering less normal tissue, but to a higher dose. This is because of three main considerations:

1. In my experience, the particular geometry of the tumor and normal tissues in a given case often allows the choice of beam directions which can advantageously spare specific organs which commonly limit the intensity of treatment which can be given, and this may be better than just spreading the dose around throughout the entire patient cross section.
2. As discussed in Chapter 5, the bath of dose around a given OAR may negatively impact its response to radiation. Thus, constraining the dose bath to a small volume seems wise.
3. Until the recently, common practice over several decades favored the few-fixed-fields approach over arc therapy – and I tend to give considerable weight to established experience. The advent of IMRT has challenged this preference, since it favors plans which cover a large fraction of the patient cross-section. However, this arises from an algorithmic need, rather than being motivated by biological considerations – and I tend to distrust changes which arise from purely technological limitations.

UNCERTAINTY IN THE DOSE DISTRIBUTION

In fairness to you, my readers, I should warn you that what I am about to say in this section reflects what I think *ought* to be done. In current practice, unfortunately, most practitioners lack the tools either to make detailed analyses of dose uncertainties, or to display the results of an uncertainty analysis. I write this in the hope that you will lobby for change.

Calculation of uncertainty

There are numerous sources of uncertainty in the estimation of the dose delivered to the patient, many of which have been touched upon in preceding chapters. Chapter 2 will surely have convinced you of the necessity of assessing, displaying, and recording uncertainties and the confidence level at which they have been estimated. If you are not yet convinced, let me give you a scenario that highlights the problem that arises if one does not analyze uncertainties. Imagine a clinician has set a dose constraint that the center of the spinal cord not receive more than 48 Gy and that the planner has developed a plan in which the dose to the center of the cord is precisely 48 Gy. Upon seeing this, an unwary clinician would probably be satisfied and would sign off on the plan provided, of course, that the other constraints were also satisfied. Now, suppose the planner was a little bit savvy about uncertainties and was to warn the clinician that, yes, the best estimate of the cord dose was indeed 48 Gy, but that there was a 50% chance that the cord dose was higher than that (as is, indeed, the case). The clinician's attitude towards the plan would almost certainly change. He or she would want to know how much over 48 Gy (at some confidence level)[5] the dose could be before agreeing to the plan. That is, he or she would want to know the upper bound on the dose estimate.

Although quite some attention has been given to specific sources of uncertainty, such as patient and organ motion as discussed in Chapter 7, there has been little done to quantify the overall uncertainty in the dose delivered throughout the patient. I have proposed a simple approach to this problem (Goitein, 1985), which involves computing three dose distributions, namely the nominal, upper-bound, and lower-bound dose distributions. The "nominal" dose distribution is based on the best estimate of all factors involved in computing the dose. The "upper-bound" dose distribution uses extreme values (at a specified confidence level – I tend to use 85% as mentioned in

[5] I once had the chance to meet the US president's science advisor who, in conversation unrelated to the reason for my visit, was bemoaning his difficulties in presenting members of congress with an analysis of the safety of a space mission being contemplated at the time. "They don't want to hear that there is only one chance in ten million of a problem. They want to know: is it safe, or not?" We in the world of radiation oncology cannot hold out for such certainty; we must learn to be comfortable with probabilities.

Chapter 2) for all the factors involved in the calculation. Thus, for the upper-bound calculation: the aperture is made larger by an amount intended to characterize patient and organ motion and registration errors; the CT densities are reduced by an amount intended to characterize the uncertainty in CT number; the dose is raised everywhere by an amount intended to characterize the possible variations in dose monitoring and calibration; and so forth. The "lower-bound" dose calculation uses the opposite extremes.

These three dose calculations allow one to quote a dose with uncertainty bounds at every point within the patient. The distributions should be interpreted with caution since the uncertainties at different points are highly correlated. As a consequence, neither the upper nor the lower-bound dose distribution is physically possible and the displays tend to overestimate the amount of tissue within which there might be a problem. Nevertheless, this approach provides a crude estimate of uncertainty that can be useful in warning of possible problems and can lead to a search for more "robust" solutions whose uncertainty bounds are smaller.

Display of uncertainty

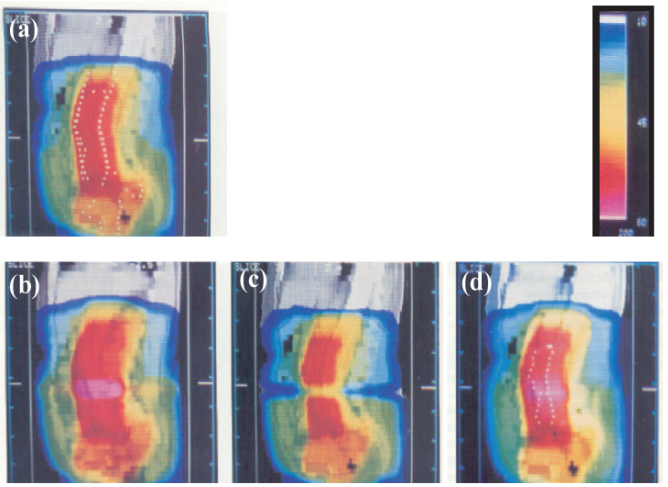


Figure 8.9. Coronal sections of a plan which, due to the large field sizes involved, required abutting superior and inferior fields. (a) Nominal dose distribution, (b) upper bound dose distribution, (c) lower bound dose distribution, and (d) upper bound dose distribution when the junction between the fields is feathered (85% confidence limits.) Reproduced with permission from Urie *et al*. (1991).

The presentation of the uncertainty in a three-dimensional dose distribution presents a challenging problem due to the plethora of data. One approach is that described in Urie *et al.* (1991) and Goitein (1985), an example of which is shown in Figure 8.9. In this case, because of the large target volume and limitations on field size, abutting fields had to be used. In Figure 8.9, three dose distributions are juxtaposed: (a) the nominal (most likely) dose, (b) the upper-bounds on the dose, and (c) the lower-bounds on the dose at each point at the stated probability level. This figure highlights the scale of potential problems that can arise at a beam juncture due to possible treatment uncertainties and, in Figure 8.9d, how the size of these uncertainties can be reduced by beam feathering.

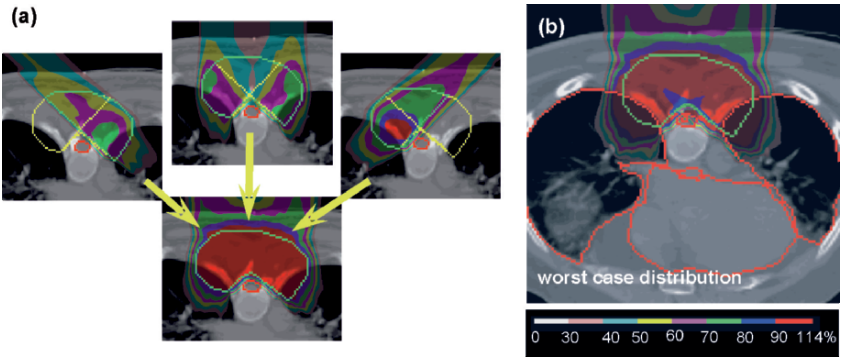


Figure 8.10. Uncertainty in a three-field plan employing proton beams. (a) The dose distributions in a transverse section for the three individual beams together with the composite (nominal) dose distribution. (b) The "worst case" dose distribution in the same section (see text). Reproduced with permission from ICRU78 (2007).

An alternative approach has been developed by Lomax (ICRU78, 2007). In this method, dose distributions are calculated for a number of translated or rotated CT data sets, and, potentially, from data sets with altered CT numbers to simulate density uncertainties. A hybrid dose distribution, that indicates the worst-case dose at any point, is then computed as follows. For points within the PTV, the dose is set to the lowest dose at that point in any of the plans. For those points outside the PTV and, hence, within normal tissue the dose is set to the highest dose in any of the plans. This one display thus shows both potential cold spots within the tumor and potential hot spots within normal tissues. The result of such an analysis is shown in Figure 8.10; the potential cool regions in the tumor, colored blue corresponding to a 10 to 20% dose reduction, are due to possible junction

problems with the three abutting beams. This presentation has its origins in the display of "images of regret" suggested by Shalev *et al*. (1988).

Uncertainty in quantities depending on the dose distribution

One is also, of course, interested in the uncertainties in quantities derived from the 3D dose distribution. This includes scalar quantities such as D_{min}, $D_{near\text{-}min}$, D_{mean}, $D_{near\text{-}max}$, D_{max}, and so forth for all VOIs, and estimates of biophysical quantities such as TCP, NTCP, and EUD. The uncertainty bounds in these quantities can be estimated (in fact, probably over-estimated for the reason given above) by computing their values from the lower- and upper-bound dose distributions. An approach to estimating the uncertainty bounds of DVHs has been presented by Niemierko and Goitein (1994).

THE PATIENT'S-EYE VIEW

I have discussed the process of planning and delivering radiation therapy as though it were exclusively the domain of the treatment planner and the patient's physician. However, the patient is a vital part of this process not just a passive recipient of the treatment. The patient has to be involved in numerous aspects of the process.

Diagnosis and choice of treatment modality

Self-evidently, the patient's self-reporting (the history) is an important element of diagnosis. The patient also can play a central role in the choice of therapeutic modality. He or she may have quite personal views on, say, organ preservation, which can sway the choice of modality between, say, surgery and radiation therapy. It is vital that the patient gives informed consent to the treatment decided upon. For this, the patient must be *fully* informed. In my years of working in a radiation therapy department, I have heard several conversations in which the plan for the patient's therapy was being presented to the patient by his or her physician. Too great a fraction of these involved more of a lecture than a give-and-take conversation.

The patient's role in risk management

The balance between risks is central to treatment planning decisions. Most notably, between local tumor control and normal tissue complication probabilities, but also among the various normal tissue complication probabilities since it is sometimes possible to spare one organ at the expense of another. The patient may be a willing

risk-taker and opt for an aggressive treatment, or may prefer a more conservative approach. And, he or she may be particularly interested in avoiding specific morbidities. These considerations are certainly in the clinician's mind in formulating the planning goals, but I believe the patient needs to be brought into the decision-making process more often and more explicitly than is the custom.

The patient as monitor of the treatment

Experienced therapists have often had the experience of patients reporting unusual events during their treatments Seemingly minor observations (unusual noises, unusual session duration, bodily reactions, etc.) may be harbingers of danger.

I had an experience, very early on in my career in radiation therapy, that has stayed vividly in my mind. I was responsible for planning a treatment for a patient for whom I had designed and laboriously hand-made a compensating filter (see Chapter 4). The patient had a pelvic tumor with a sloping lower torso and the compensator was designed to deliver a uniform dose at the depth of the tumor. The treatment machine was a ceiling-mounted 2 MeV van der Graaf accelerator whose beam was pointed downwards at the patient, lying on a couch. After the his first treatment session, I was informed that the patient wanted to talk with the person who was responsible for the technical details of his treatment. Assuming he wanted to congratulate me on my fine work, I hurried round to see him. "As I lie down on the couch" said he, "I can see the gadget that you made for me, hanging below the machine. Doesn't that mean that scattered radiation from it is reaching my eyes? What is the dose to my lens?" (See Figure 8.11). I was enormously impressed by his acute observation and common-sense. I had no idea what the answer was, and spent that evening measuring the scattered dose – which turned out to be acceptably low, but by no means negligible.

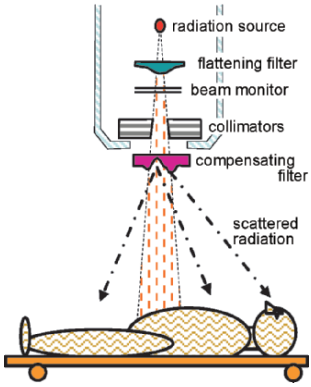


Figure 8.11. Schematic view of a patient beneath a treatment machine. The patient could see the compensating filter, from which he deduced that scattered radiation could reach his eyes.

From that day on I have regarded patients as technical partners in my work.

9. IMRT and "OPTIMIZATION"

INTRODUCTION

I have alluded to intensity-modulated radiation therapy (IMRT) several times, and there is a brief introduction to it in the introductory Chapter 1, but now it is time to discuss it in greater depth.

The concept behind IMRT is that, in order to deliver some desired (not necessarily uniform) dose distribution throughout the PTV, the

fluence within any one beam need not be uniform across the PTV. The non-uniformity of each beam is driven by anatomic features specific to the patient and allows a better sparing of nearby normal tissues than is possible with uniform-intensity radiation therapy. IMRT was developed independently by Cormack (1987) and Brahme (1988)[1]. IMRT has become widely accepted as a worthwhile approach for patient treatments. Palta *et al*. (2003) and Bortfeld *et al.* (2006) offer extensive accounts of IMRT, and Bortfeld (2006) provides a short overview of IMRT with references to many of the important papers.

IMRT allows one to deliver only low doses to all or part of selected normal structures. In particular, one can create a concave irradiated volume that can spare much of an invaginating OAR, whereas uniform-intensity radiation therapy inherently creates convex irradiated volumes and cannot achieve such OAR sparing. This difference is illustrated in Figure 9.1.

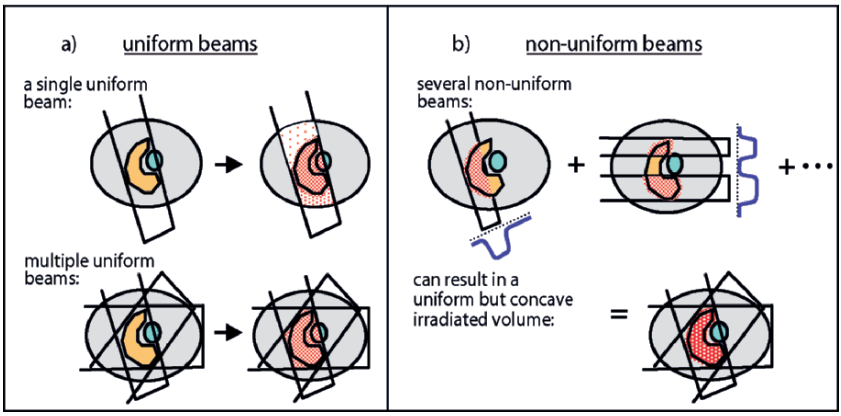


Figure 9.1. Illustration of why IMRT, but not uniform-intensity radiation therapy, can generate concave dose distributions (see text).

Uniform-intensity radiation therapy is illustrated in the left hand side panel. A beam that covers the target volume with a uniform flux of radiation cannot avoid irradiating an OAR invaginating the target volume. This is true of each beam. The irradiated volume is the intersection of the tissues irradiated to high dose in each of the beams.

[1] Pedroni (1981) had earlier used intensity modulation in the context of π-meson therapy.

Since the volume irradiated to high dose in each beam is convex, their intersection must be convex and the part of the OAR within that volume will inevitably receive full dose.

IMRT is, rather simplistically, illustrated in the right hand side panel. Here, each beam can be blocked so as to avoid the invaginating OAR. If this blocking is done, the target volume will not receive dose in the shadow of the OAR, and so will have an inherently inhomogeneous dose distribution from that beam. However, other beams, coming from other directions, can "fill in" the dose that the first beam failed to deliver. Consequently, through the use of several non-uniform beams, a fairly uniform target volume dose can be achieved, while largely sparing the OAR. This approach can thus create concave dose distributions. The strategy of sparing selected normal tissues has been given the name *conformal avoidance*, in analogy with the traditional conformal coverage of target volumes.

It is not only invaginating OARs that can be spared radiation; other selected neighboring or distant OARs can also be spared or partially spared. On the other hand, it is not possible to spare *all* the OARs – the integral dose has to be deposited somewhere – one can only spare some few selected ones. In general, the smaller the OAR, the easier it is to spare it.

IMRT also makes it simple to deliver a non-uniform dose distribution to the target volume. In general there are two situations in which a non-uniform dose distribution may be desired. First, when there are two target volumes, one nested inside the other, and one wishes to deliver a higher dose to the inner volume than the outer, all in the same fraction (the so-called "field-within-a-field" approach). This might be the case when the inner volume encompasses only the GTV and the outer includes sub-clinical disease. The second situation is in the case of so-called dose-painting (Ling *et al*., 2000). Dose painting may involve delivering additional dose to sub-regions of the target volume due to the judgment that, based perhaps on functional imaging studies, they contain more resistant cells, Or, dose painting may be desired in order to deliver a reduced dose to sub-regions of the target volume because a critical normal tissue runs through, or is closely adjacent to, the target volume.

Figure 9.2 shows an example of an IMRT plan. A transverse section of a patient with a nasopharyngeal carcinoma is shown, irradiated by nine equally spaced photon beams. The dose distributions of the individual beams are shown around the periphery of the central larger

image, which shows the dose from all of the beams combined together. One sees vividly how the design process has led to the individual beams having greatly reduced intensity where they shadow the centrally located brain stem, while the target volume has, nevertheless, been covered quite uniformly.

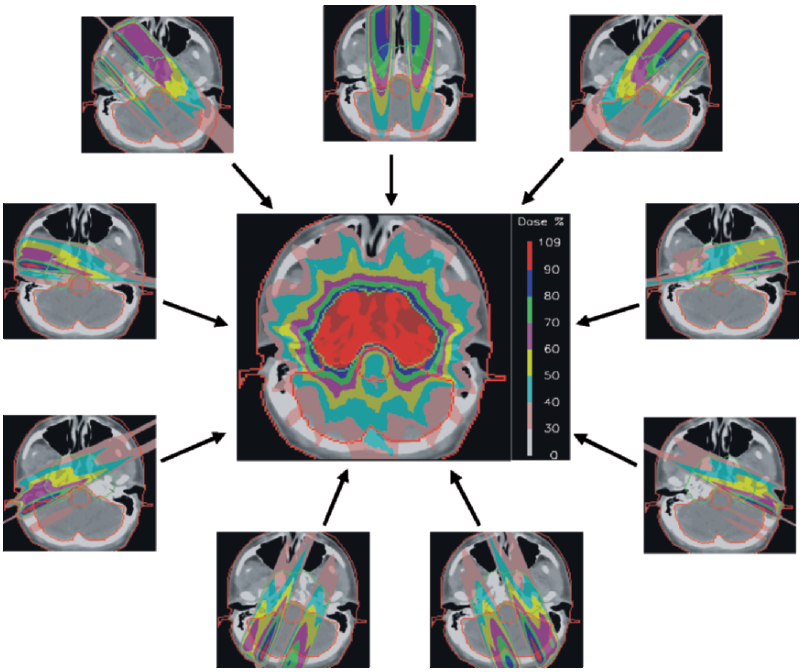


Figure 9.2. IMRT of a nasopharyngeal carcinoma using nine photon beams, equally spaced in angle. The dose distributions of the individual beams are shown in the surrounding panels, and the overall distribution in the central panel. Figure courtesy of A. Lomax, PSI, CH.

THE DEVELOPMENT OF AN IMRT PLAN

How does one go about designing an IMRT plan?

Inverse planning of IMRT

The original idea was that one could use an analytic process that, given a desired dose distribution, would determine the treatment

variables needed to achieve it. The mathematics for this approach was developed by analogy with the process of CT reconstruction. The flow chart for this process is shown in Figure 9.3.

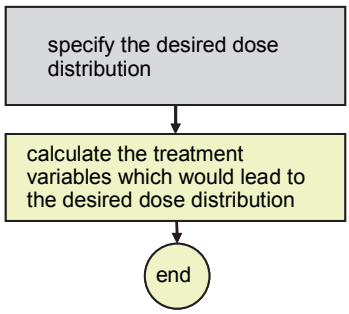


Figure 9.3. Flow chart of the process of inverse planning.

This beguilingly straightforward scheme has, alas, some deep problems associated with it:

1. Initially, the desired dose distribution was specified as a distribution whose value was the desired tumor dose within the target volume, and zero everywhere outside it. That would be ideal. However, the laws of physics imply that such a distribution is physically unrealizable. When the mathematics of inverse planning was applied to the ideal dose distribution, it returned physically unachievable values for some of the treatment variables. Namely, it required the use of beams in parts of which the intensity was negative. A negative intensity beam would be one that would suck dose out of the patient. How nice, if it were possible.
2. The solution to the first problem was to reset the negative intensities to zero, and live with the dose distribution that then resulted. This solution indeed led to some very interesting concave dose distributions which demonstrated the great potential of intensity modulation. However, there was no room for balancing conflicting goals such as tumor control and morbidity of normal tissues. This balance is central to radiotherapy and the planning of radiation treatments and the lack of a way to effect that balance made this approach unattractive.
3. If one could define a desirable dose distribution that met all one's goals and constraints *and was physically realizable*, then inverse planning would be the perfect way of getting values for the treatment variables. However, this is a circular argument. One does not know what such a dose distribution looks like and it is an impossible task to predetermine it before having performed the treatment planning process. If you doubt this, just try it.

For all the above reasons, inverse planning has not been successfully applied in clinical practice. What is done, very successfully, is "forward planning" of IMRT, the process to which we now turn our attention.

Forward planning of IMRT

The process of forward planning is shown in Figure 9.4. It will not escape the alert reader's eye that this flow chart is virtually identical to Figure 8.3 of the previous chapter, in which the process of manual planning was illustrated. The only differences are the following.

(1) The treatment variables, outlined in cyan, now include the fluence maps of each beam, and so there are very many more variables than is the case for uniform-intensity beams.

(2) Many, but not all, of the starting values of the treatment variables, being so numerous, are selected by the computer.

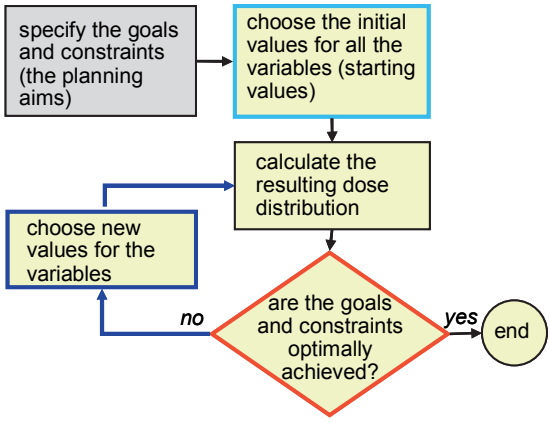


Figure 9.4. Flow chart for the forward planning of IMRT.

(3) The evaluation process, outlined in red, is now performed by the computer, rather than the human planner.

(4) The iteration loop, outlined in blue, is performed by the computer, which decides on the changes in the variables for the next iteration and performs the iterations up to hundreds or even thousands of times rather than just a few times as is the case in manual planning.

Because the origins of IMRT were based on the use of inverse planning, that term has become widely used to describe the process of designing an intensity-modulated radiation therapy plan. However, this is a misnomer. When one designs an IMRT plan, one is using forward planning.

Planning IMRT

There are two main aspects of planning IMRT:

1. establishing a method for computing a numerical *score*, expressing how well the goals were achieved; and
2. conducting a *search* through the space of treatment variables to locate the set of values of those variables that gives the best score.

The score is the value taken on by a *score function* for a particular set of the variables upon which the value of the function depends.[2] What one is attempting to do is to pick values for all of the variables that together maximize the score.

In planning IMRT, the processes of establishing a score function and searching for the optimum score are set within a broader range of activities, namely:

1. Evaluate the patient using all relevant diagnostic tools, and decide whether to employ radiation therapy as at least a part of the patient's treatment.
2. *Obtain* and inter-register appropriate imaging studies. The planning CT study, which is taken with the patient lying in the position and, usually, held in the immobilization device that will be used for treatment, is almost always one of these studies.
3. Delineate on the planning CT the target volumes (GTV, CTV, and PTV) and all OARs (and, perhaps, PRVs) whose proximity to the target volume or sensitivity makes them of particular interest.
4. Establish the planning aims for the treatment.
5. Set or change values for the treatment variables. This defines a plan – namely, a set of beams with, in general, non-uniform fluence maps together with the beam weights.
6. Evaluate the plan (i.e., compute its score) and either select it for use in treatment or continue the search by adjusting the values of the treatment variables and returning to step 5.
7. Finalize the prescription.

[2] The more formal term for this function is the *objective function.* However, I use the synonymous term *score function* in what follows as being more evocative.

8. Simulate the selected plan to ensure that it is deliverable and that all parameters have been correctly established.
9. Deliver the treatment, and verify that the delivery is correct, in many fractions over many weeks.
10. Re-evaluate the patient during the course of treatment to ensure that the plan remains appropriate (e.g., weight loss or tumor regression have not affected the treatment geometry unduly) and, if it does not, return to step 5, or even 2, to re-plan the remainder the treatment.
11. Document and archive the final treatment plan.
12. Review the treatment plan at the time of patient follow-up or possible recurrence.

I have used up a lot of paper and ink here to emphasize one simple but important point, namely, that *these steps are identical to the planning steps outlined in Chapter 6 with the exception of steps 5 and 6*. That is, there is a great deal of work which is common to manual planning and computer-driven planning. I will confine myself for the rest of this chapter to discussing those aspects that are unique to the latter.

There are numerous tried and true mathematical techniques which are designed to search in a large space of variables for the set of values of those variables that maximizes the value of some function of those variables; that is, which maximizes the score. Since the search for an IMRT plan uses such so-called optimization techniques, it is often referred to as a process of *optimization*. As I shall explain at the end of this chapter, I don't like this term. Nevertheless, I bow to common usage and employ it here.

The following is a breakdown of steps 5 and 6 in the preceding list. The italicized items are ones that require human input:

a) Design those aspects of the plan that you do not plan to optimize– for example, choose the beam directions.

b) Establish the goals of the optimization process (deduced from the planning aims) – for example, the score function you wish to optimize, and the constraints and their importance factors, if any.

c) Set any parameters needed by the search algorithm, if any.

d) Provide starting values for all variables that will be set in the search process – many search algorithms do this for you, automatically and invisibly.

e) Perform an iterative search for an optimal solution.

f) Evaluate the "optimal" solution found and, if unsatisfactory, modify the constraints, importance factors and so forth and repeat the search process until satisfied.

Both the development of a clinically meaningful score and the search process to optimize it will shortly be discussed. First, however, I want to point out the size of the problem.

Magnitude of the optimazation problem

It is informative to estimate the number of values of the treatment variables that need to be explored in picking the optimal therapy. Table 9.1 gives a sense of the magnitude of the problem. The numbers in this table are based on estimates of what constitutes a significant change in each variable. For example, it is assumed that there are approximately 50 distinguishable settings for the gantry angle. That is, that a difference of about 7° in a beam's angle would just be significant.

Table 9.1. A partial list of the variables upon which a radiation treatment may depend, together with a rough estimate of the number of significantly different values the variables may assume.

	the typical number of variables	*approximate number of significantly different values of the variable*	*the overall number of significantly different choices*
THE "GIVENS"			
the total dose	1		
the dose constraints	10		
THE TREATMENT FIELDS			
the number of fields (2 -> 20)	6	1	6
for each field:			
the direction of each field relative to the tumor	5	50	250
the field size (+x, -x, +y, -y)	4	50	200
the field shape (~ 20 points)	40	50	2000
the overall field weight	1	5	5
weights of pencil beams within a field (IMRT)	1000	5	5000
OVERALL (6 field treatment)			
3D-CRT (without IMRT)	~ 300		3.E+09
3D-CRT with IMRT)	~ 6'300		1.5E+13

Even without intensity modulation within each beam, there are some $3 \cdot 10^9$ distinguishable possibilities. If it took only a millisecond to evaluate each one, a computer would take over a month to assess all possibilities and choose the best. In the case of IMRT, the variables

which characterize the non-uniform intensity of each beam constitute the *fluence map* of the beam, which is characterized by an array of at least 30 x 30 intensity values. Thus, in IMRT, there are far more distinguishable possibilities – some 10^{13} *in toto* – and it would take about 1,000 years to assess all possibilities and choose the best. In light of these numbers, *evaluation of all distinguishable plans is not feasible*. Instead, one must resort to intelligent search algorithms that examine only a very small subset of all possible plans.

It is worth noting that optimization is potentially useful in uniform-intensity radiation therapy as well as in IMRT (Niemierko, 1992). There are only two related differences: optimization of uniform-intensity radiation therapy requires the setting of far fewer variables than does IMRT; and, for that reason, optimization is essential for planning IMRT, but not for uniform-intensity radiation therapy.

Scoring vs. Searching

In the past, the attention of those attempting optimization was focused on the search process. This was for a couple of reasons. On the one hand, the scale of the problem makes it mathematically challenging and, therefore, technically interesting. On the other hand, the problem of searching in a large space of possible solutions to find the optimal solution is one that has received a great deal of attention from mathematicians, and there exists a shelf-full of mathematical approaches which can be taken over and tried. However, *no matter how hard it may be to develop an efficient and reliable search process, it is much harder to develop a realistic and reliable measure of plan goodness. In consequence, it is the latter problem which should receive the lion's share of attention*.

Let us take a peek at our planner as he develops an IMRT plan using the optimization scheme illustrated in Figure 9.4. Figure 9.5 is a sneak view of him. He has his hands folded and appears to have nothing to do except to stare at the meter

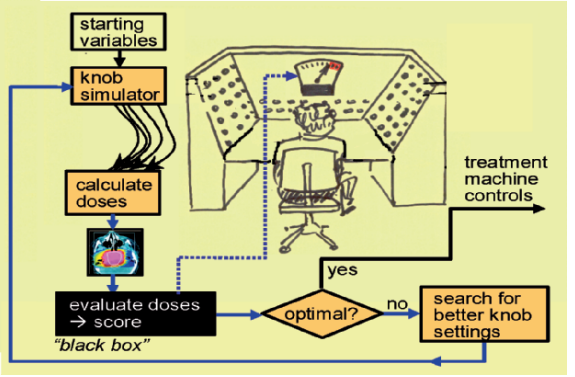


Figure 9.5. IMRT being planned

which shows the current score steadily creeping up as the plan is iteratively improved. Of course, this is an oratorical exaggeration. There will still be plenty for our planner to do, as we will shortly discuss.

WHY SCORE?

It is rather disconcerting to have to boil down one's impressions of a treatment plan, with all the complex considerations that go into evaluating it, into a single value, the numerical score. I first discuss why one has to do this – and then, how one goes about doing it.

Plan comparison

In manual planning, the planner in Figure 8.2 can generate a handful of plans – but a computer can outstrip him by orders of magnitude. However, in the end, either the planner or the computer is going to have to choose one and only one plan for the patient's treatment. Undoubtedly, one will wish to choose the best plan that one has come up with. This means that, perforce, one has to be able to make the judgment of which of two plans is the better. The formal way of saying that is that one has to be able to "rank" the two plans. What if one has generated not two, but several plans? One will surely want to use the best of them all. Well, if one knows how to rank any two plans, then one can rank all pairs of the plans one is considering and thereby find the best.

Unfortunately, there is a flaw in this argument. Suppose our planner was a bit distracted and, looking at all pair-wise plan comparisons, he decided that: plan B was better than plan A; plan C was better than plan B; and that plan A was better than plan C. Well, he suddenly has a logical conundrum on his hands; the three statements are incompatible with one another. This incompatibility brings to light the fact that merely ranking plans, given that one can be inconsistent, may not be sufficient to choose the best of them. What is one to do?

Plan quantitation

The answer to the preceding question lies in another aspect of ranking. If one says that one plan is better than another, then one is somehow taking a large number of considerations and rolling them

into one overall idea of "goodness." Only then can one say that one is better than another. An unambiguous way to rank objects is to assign a numerical value to each of them. That is, to combine all their aspects into a single judgment and express that judgment in the form of a numerical "score." The plan with the highest score will then be, in the scorer's judgment, the best plan. Scoring has another advantage. In addition to averting the logical impasse which ranking alone can get one into, it also allows one to express the idea that two plans are "not very different." If their scores are quite similar, then presumably they will be of quite similar goodness. One could say that our logical problem of inconsistent rankings came about because of uncertainties in scoring.

In manual treatment planning, quantitation of a score is not much needed and is rarely, if ever, attempted. Rather, the planner is likely to make pair-wise comparisons and live with the logical difficulties that can result. He will understand that, when two plans are very close, it is not critical which is ranked higher. However, in computer-based optimization, quantitation is necessary. A computer can only rank plans by giving each a score and then selecting the plan with the highest score.

WHAT IS OFTEN NOT IN THE SCORE

The optimization schemes used to date largely restrict themselves to calculating the overall weights and fluence maps of each defined beam. The beams themselves must, at the current level of the technology, be pre-designed by the planner as regards to modality, direction, and number – although some optimization schemes can also constrain the total number of beams used from among a larger number of pre-selected beams. So, the so-called optimization of radiotherapy only optimizes a subset of the treatment variables. The others have to be determined by the planner. The following discusses some of the beam properties that are generally not included in the optimization process – or, expressed more precisely, are not included as variables in the score function.

Modality

The choice of modality (photons, electrons, protons, etc.) is usually based on the planner's experience – and can be different for each

beam. However, it is possible in principle to have both the beam modality and energy as variables in the optimization process. This can be done, for example, by including beams of different modalities in the set of optimized beams, and then keeping only the "best" beams.

The number and directions of beams

The choice of the number and direction of beams is a very important one. Indeed, in manual planning, these are among the most important of the planner's decisions, strongly affecting, as they do, the extent of coverage of the normal tissues outside the target volume. The fact that current optimization schemes generally do not attempt to adjust these variables is a serious weakness.

As regards direction, it is common in automated optimization to pick the beams to be coplanar and equally spaced in angle, all around the patient. This reflects a tendency for intensity-modulated radiation therapy to prefer isotropic beam arrangements since the ability of one beam to compensate for a dose deficit in the target volume left by another beam generally requires that the beams be well separated in angle from one another. The restriction to coplanarity is entirely unnecessary and reflects limitations either in the planner's imagination in the planning system's capabilities, rather than deficiencies in the mathematics of optimization.

There are efforts underway to develop criteria for choosing the beam directions more intelligently. These, for example, may try to take advantage of geometrical features of the patient's anatomy to determine some preferred beam directions, as is done in manual planning. However, these efforts have not yet reached routine practice.

The choice of the number of beams to use is a matter of some controversy. There are proponents of the value of having a large number of beams – in the limit, a full 360° beam delivery. Such capabilities are being developed under the name of Tomotherapy (Mackie *et al.*, 1993). The mathematics of tomographic reconstruction, indeed, suggests that the use of beams covering the full 360° would provide the highest conformity of dose to the desired prescription. However, there are strong proponents of a much more limited number of beams. At present, a number of around seven equally spaced beams is commonly used. It is usual to pick an odd number of beams, since having pairs of parallel-opposed beams

are "wasteful" in that they cannot do much more than a single beam in terms of conformal avoidance. There is some suggestion that fewer beams are needed when heavy charged particles such as protons are employed.

Lateral extent of beams

There is little reason, at first glance, to have any beam intensity for those pencils which are not directed toward the target volume since they appear only to intersect normal tissues and not the target. For this reason, such pencils are usually set to zero intensity and excluded from further optimization. Doing so, however, can lead to problems with dose at the edge of the target volume. The reason is that the pencils directed toward the target volume, but close to its edge, lose dose through electron transport and angular beam divergence to tissues outside the target periphery. In a uniform intensity beam, designed to cover the target volume *with some margins*, this loss of dose is exactly compensated for by pencils directed just outside the target volume which contribute dose to the parts of the target just inside its periphery. When the outer pencils are lacking, there is then a dose reduction at the target edge.

Many algorithms use some kind of trick to prevent such a dose reduction. For example, the intensities of those pencils that are directed towards points some small distance outside the target volume can be included in the optimization. That distance then becomes a parameter of the algorithm. A user needs to be aware of such often-hidden algorithmic features.

THE SCORE

The state of our knowledge of the impact of radiation on normal tissues is so inadequate that it casts considerable doubt on the realism of present techniques for assigning a score to a plan. As discussed in Chapter 8, we still do not know the answer to the very basic question: is a 4-field box in better or worse than a 360° rotation in a given situation? If one cannot answer even this question, then one can hardly expect to be able to answer many of the other questions of concern in planning – and optimizing – treatments.

Nevertheless, to perform any type of optimization, one must compute the best score one knows how to produce - just as, to treat a patient, one can only base the plan on one's best judgment. This score should combine elements, direct or indirect, that provide measures of:

- the likelihood of local tumor control;
- the likelihood of morbidity;
- other aspects of the plan such as its complexity and feasibility. (This important aspect of a plan, which planners take into account subliminally when planning by hand, is generally not considered in IMRT planning.)

It will not escape your notice that *these issues are exactly the same issues that a planner faces during manual development of a plan*. The difference is that, while the planner does the analysis in his head, the computer must do the analysis by computation.

Unfortunately, many score functions are designed more for computational convenience than for clinical appropriateness. The deficiencies fall into two classes: (1) The parameters being optimized are too crude – they may not include measures of dose correlated with normal tissue volume although, as I have tried to emphasize, dose-volume effects are very important. Or, (2) they are of a form (linear or quadratic in the variables, for example) which simplifies or speeds up the search process but has no medical basis. It is my opinion that the reason that early attempts at optimization in the 1960s and 1970s largely failed was because their score functions had almost no clinical grounding and took no account of dose-volume effects.

Measures that describe a plan's impact

The following quantities, either singly or in some combination, measure the impact of a plan on the patient. (*N.B.*, the score is, of course, based on the total overall dose, not the dose per beam.)

Tumor response

- the difference between the minimum (or mean, or…) target volume dose and the prescribed dose
- the dose received or exceeded by 95% of the target volume ($D_{95\%}$), which reflects the depth of any cold spots
- the dose received or exceeded in 5% of the target volume ($D_{5\%}$), which reflects the height of any hot spots

- the degree of dose homogeneity within the target volume, sometimes expressed as the root-mean-square of the differences between the dose at each voxel of the tumor and the mean target dose
- the estimated tumor control probability (TCP)
- the estimated equivalent uniform dose (EUD)

Morbidity (for each organ at risk)

- the difference between the maximum (or mean, or…) OAR dose and its constraint dose;
- the difference between the volume (relative or absolute) of the OAR which receives a dose of D Gy or more (V_D) and the corresponding dose-volume constraint. There may be several such requirements for a single OAR;
- the estimated normal tissue complication probability (NTCP) for the endpoint(s) of interest;
- the estimated equivalent uniform dose (EUD);
- the integral dose delivered outside the CTV.

Complexity and Feasibility

- number of beams
- use of unusual beam directions (e.g., non-coplanar beams)
- the need for unusual patient positioning or immobilization

The score is computed from one or a combination of the above measures, suitably weighted. It is a single number. The computation of the values of these measures is straightforward and is readily performed by the computer, except for some hard-to-quantify aspects of plan complexity and feasibility. If only one measure is maximized, such as the mean target volume dose, there is no further problem. However, if more than one of these measures are combined into a score, the problem becomes immediately much more complex. The way in which the measures should be combined and, in particular, the assignment of an importance factor to each measure are matters which must be decided by clinical and often subjective means.

Use of biophysical models

It is partly because of the arbitrary nature of the dose-based measures identified above, and their almost unknowable relative weighting factors,[3] that interest has grown in the use of biological models such as TCP, NTCP, and EUD for optimization. While confidence in the estimation of these quantities may be poor, this is compensated by the fact that *they have an intuitively obvious clinical meaning*. As a result, the relative importance of a given increase in one of them, say the TCP, and in another, say the NTCP for pneumonitis, can be readily understood in human terms. Indeed, these are quantities that the patient too can understand and the relative weighting of which he or she may have an opinion about which should be taken into account.

Use of a score in optimization

There are two ways to use a score in the process of optimization:

Score optimization In this approach all the variables to be determined are bundled into the score function and the optimizer must choose the values of those variables that maximize the score. Score optimization is what is done when, for example, the probability of uncomplicated control is (mis)used as the score function (see below).

Constrained Optimization The second approach is one in which the optimizer seeks to maximize the score *subject to constraints on one or more measures* such as those listed above. These constraints form a threshold above or below which a given measure must lie.[4] As an example of *constrained optimization*, one might seek to maximize the mean dose to the target volume subject to the requirement that the maximum dose to each organ-at-risk of interest is less than the predefined maximum allowed dose for that organ.

---

[3] It is very hard to estimate numerically, for example, the relative importance of the mean dose to the target volume and the standard deviation of dose (a measure of dose inhomogeneity) within the target volume. Or, how can one know the relative importance of the $D_{95\%}$ of the target volume and the V_{20Gy} of an OAR?

[4] This was the approach reported by Niemierko (1992) in which the user was allowed to pick both the measure to be used as the score function and those to be used as constraints, from a long menu of possibilities.

In constrained optimization, as just described, the constraints would be so-called "hard" constraints in which *no violation of the constraint whatsoever is allowed*. It is also possible, and more realistic to define "soft" constraints – for which a gradually increasing penalty is levied the further one gets away from the constraint value. Soft constraints are helpful in many search processes as they avoid some technical problems, and are much closer to clinical intentions – it is rare that a constraint has to be met exactly, there is usually the possibility of accommodating small violations of it. To allow minor constraint violations, one must be able to compute a penalty for the constraints to prevent the violations from being too great. "Too great" must be related to the value of the one thing being optimized – so that one cannot escape the difficult problem, already encountered in the discussion of score optimization, of including factors that relate the level of a constraint violation to the improvement in the score.

Of these two approaches to optimization, the second approach, that of constrained optimization, seems to me to lie much closer to the way planners assess plans during the process of manual planning, and I favor its use.

Estimating the tumor response

In planning as presently performed, the tumor response is taken to be determined by the overall dose distribution delivered to the tumor; other factors such as fractionation are usually not included in arriving at a score. When one wishes to deliver a quite uniform dose to the PTV, the mean tumor dose or the $D_{50\%}$ or the $D_{98\%}$, for example, can be employed as indirect measures of the implication of a plan for the TCP.[5]

Other possible measures of tumor response are: the EUD, which tries to take the dose heterogeneity within the target volume into account in a quantitative way; and the TCP, computed with the help of a biophysical model.

In addition to a direct measure of tumor response, one often places constraints on two additional factors. The first is the dose received by the normal tissue(s) within the target volume. The tumor may be

[5] Provided that the dose heterogeneity within the target volume is small; for example, that the standard deviation of the dose variation within the target volume is less than some defined percentage of the mean dose.

interspersed within normal tissue, or may be supported on normal tissue stroma, and the PTV certainly includes normal tissues outside the CTV. To avoid damage to these normal tissues, one often places an upper bound on the dose within the PTV. The second factor is the dose inhomogeneity. Even if one is pretty confident in the model used to estimate TCP or EUD, one may not trust it enough to allow it to accept a highly inhomogeneous dose distribution within the target volume. Thus, it is common to place a dose constraint on the dose inhomogeneity. This constraint can be achieved by placing upper and lower bounds on the dose within the target volume, or by placing a constraint on the difference between the EUD and, say, D_{mean}.

Estimating normal tissue response

As discussed in Chapter 6, in manual plan evaluation one tends to separately inspect each organ at risk (OAR) and the remaining volume at risk (RVR) to assess the impact of the dose distribution delivered to each of those volumes of interest. The same approach is fully appropriate to, and widely used in, computer-driven planning. For each OAR, one computes a quantity such as V_{20Gy}, or EUD, or D_{max}, or NTCP, and assigns a sub-score to that OAR based on the computed value. The sub-score is then either used in computing the value of the score function, or is used in connection with a dose or dose–volume constraint.

In practice, it is often the case that the normal tissue constraints do not adequately force the optimization process to produce acceptable results. For example, the dose may not fall off as rapidly as desirable outside the target volume because the identified normal tissue constraints are not strong enough, or are not defined for all tissues outside the target volume. De Neve *et al*. (2006) have given a good discussion of many of the tricks that can be used to achieve satisfactory results. These include defining a shell or shells around the target volume within which additional constraints can be applied, and defining a "virtual" normal tissue in a region in which a hot spot occurs which will drive the dose down in that region when the optimization process is repeated.

Combining tumor and normal tissue responses

In constrained optimization with hard constraints, one does not have to combine different elements; one optimizes one element, and places constraints on all others of interest. Score optimization is different;

one is trying to combine several elements into a single number. When the sub-scores are physical parameters such as measures of dose or volume, it is very hard to arrive at appropriate importance factors for each of them. As already pointed out, it is much easier to do so when biophysical quantities, such as TCP and NTCP, are involved. Overall morbidity can then be represented by the combined impact of the NTCPs of all the irradiated organs and tissues. From the point of view of probability theory, the "un-complication probabilities (equal to one minus the complication probabilities) are multiplicative. Therefore, one might hope to calculate the overall NTCP as in the following formula.

$$NTCP_{overall} = 1 - [\,(1-NTCP_1)\cdot(1-NTCP_2)\cdot \ldots\,]$$

However, this approach is far too simplistic. The NTCP for a given end point is in general a function of any predisposing conditions. Age, diabetes, or a history of tobacco and alcohol abuse are well-known examples of such predisposing conditions. Less well-known, unfortunately, is the quantitative impact of these conditions on the various NTCP's. Then, too, any given NTCP is defined in terms of a specific end-point. The same organ can, and indeed almost certainly will, respond in more than one way to irradiation (e.g., early and late reaction) – and, therefore, have more than one endpoint. These endpoints will be of varying gravity.

The process just mentioned of multiplying un-complication probabilities to arrive at an overall un-complication probability, *weights all complications equally*. However, a particular complication in one compartment (say, the skin) may, and in this example, certainly will, be of quite different importance than a complication in another compartment (say, the spinal cord). The uncomplication probability is, therefore, *an entirely unrealistic measure of morbidity*. To get a more realistic measure, *each complication needs at the very least to be weighted by an importance factor*. It is by no means easy to arrive at such weighting factors.

Then, there remains the problem of combining the tumor sub-score (say, TCP) with the normal tissue sub-scores (say, $NTCP_{overall}$). One approach, often cited, is that of maximizing a quantity termed "uncomplicated control" which is computed as:

$$TCP_{uncomplicated} = TCP \cdot (1-NTCP_{overall})$$

The idea behind this equation is that the goal of radiation therapy is to maximize the probability of local control of the tumor subject to there

being no complications. This approach treats tumor control and normal tissue complications on an equal footing. That is, it implies that an increase of a given percentage in a given complication can be exactly offset by an increase in TCP of the same size. But only a mathematician could accept a 5% increase in the likelihood of paralysis in order to obtain just a bit over a 5% increase in the probability of tumor control. No doctor, and no patient, is likely to agree.

In my view, the use of the probability of uncomplicated control in any optimization process is simply clinically wrong.

The patient's-eye view

So far as the patient's desires are concerned, his or her attitude toward the likelihood and nature of possible complications in relationship to the likelihood of tumor control should be taken into account – in IMRT planning, just as in uniform-beam planning. This is not just a matter of adjusting the dose delivered to tailor the treatment to the patient's degree of aversion to risk in general. As discussed in Chapter 8, the patient may have very specific concerns, the preservation of reproductive function for example, that can strongly affect the choice of beam directions and of the importance factors and, indeed, that of treatment modality.

THE SEARCH

Having decided on which variables are to be considered, and on the scoring scheme to be used, one must embark on a search to find the set of values of those variables that, together, yield the highest possible score. We turn now to this problem. If the technical details of the search process tend to make your eyes glaze over, I encourage you, rather than moving ahead a chapter or, worse still, closing the book, to jump to the final section of this Chapter, entitled "Optimization?"

The search landscape

In IMRT, one wishes to optimize a huge number of variables. The scale of the problem was suggested in Table 9.1 above. Only considering the fluence maps of each beam, the intensity profile of each beam will be divided into at least $30 \cdot 30$ pencil beams, each

with its own weight, and there are likely to be 5 or more beams. That means that IMRT requires many thousands of additional variables, over and above those needed for uniform-beam radiation therapy. All together, these form a vast hyper-space of treatment variables and the both the score and the constraints are functions of all of those variables

If only one variable were to be optimized, one could plot the score on the ordinate versus the value of that variable on the abscissa of a two-dimensional graph such as that portrayed in Figure 9.6 below, and search for the lowest point. If two variables were to be optimized, the score could be represented as a surface in a three-dimensional perspective plot such as that shown in Figure 9.7 below. The score would be represented as a sort of "landscape" with hills and valleys in it, within which one wants to find the lowest point in the deepest valley. But, we have no ability to portray a function of thousands or more variables graphically. Nevertheless, we can speak conceptually of the search landscape as a hyper-dimensional world.

How can one hope to have any possibility of success, given the vastness of the hyperspace which must be searched? That there is hope is due to several reasons. First, one virtually always selects only a subset of the variables for optimization, while fixing others such as the modality, number, direction, and shape of the beams beforehand, thereby reducing the dimensionality of the hyper-space that must be searched. Second, especially when biophysical quantities such as TCP and NTCP are used in the score function, or when particular choices are made about what parameters to optimize, the score function varies quite smoothly throughout the search space. For the most part it doesn't jump wildly around, so one may not need to look at closely spaced points. The third reason for the possibility of success is that, though it is very unlikely that one will in fact find a global extremum in a finite time, one may well find a good solution. In a sense, the possibility of success comes from the acceptability of failure.

The search itself

There is a vast literature on the subject of maximization or minimization. There are many very different, and all fascinating, methods that have been developed. And, as you might expect, there are often variants of a given method. Here I will only address two types of search techniques, without in any way giving a full

mathematical treatment. The two search techniques are "direction set optimization" and "simulated annealing."[6]

Direction Set Optimization

The *method of steepest descent* and the *conjugate gradient method* are both examples of direction set optimization algorithms. While the two methods are conceptually similar, the conjugate gradient method is the more efficient of the two and is the method that is most often used. However, I focus here on the method of steepest descent because it is more straightforward to describe, and offers a good introduction to this class of algorithms.

Consider first the one-dimensional problem in which one seeks the extremum of a function of only one variable. The extremum may be a maximum or a minimum. It doesn't matter which, since finding the maximum of a function f(**v**) is identical to finding the minimum of –f(**v**). In discussing iterative search techniques in this section, I usually talk of minimizing the score, as the graphical representations of the search process seem to me to be a bit more intuitive in that case.

The problem is represented graphically in Figure 9.6. Here, we see a landscape within which the searcher, starting at some arbitrary point S, wishes to find the minimum – but can only see a short distance around where he is. It is intuitively clear that he will be able to succeed. He could start by taking two short steps, one to the left and one to the right, and then determine which direction led to the lower score value. He would then start off along that direction. If he takes short steps until he just begins to go upwards again, he will have found the

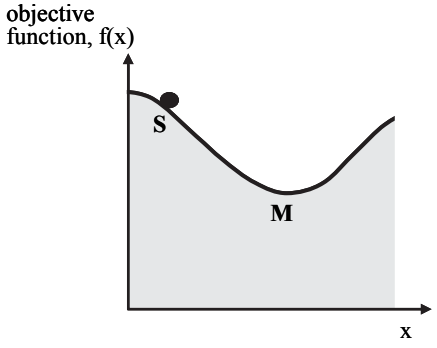


Figure 9.6: Schematic representation of the one-dimensional optimization problem (see text). x is a single variable.

[6] To learn more about search techniques, one cannot do better than read Chapter 10 of Press *et al*. (1988). This book is noteworthy for the clarity, efficiency, and unpretentiousness of its discussions.

minimum, M, to within a step's length. (There are, of course, much more efficient ways for him to find the minimum than always taking equal-sized steps.) The point is, if the searcher starts off *down* the slope, he is headed in the right direction and will eventually reach the bottom.

The direction set methods of optimization extend the one-dimensional idea to landscapes of many dimensions. They differ mainly in the way in which they pick the downhill direction, realizing that we now mean downhill in a multi-dimensional world. Figure 9.7 shows how this might look in a two-dimensional world. The searcher starts as usual at the point S. In the method of steepest descent he determines the direction of steepest descent from that point - that is, the direction in which the gradient is largest, shown by the short arrow at S. He then sets off in that direction – staying, therefore, in the colored plane of Figure 9.7. He keeps going in that plane until he reaches a minimum, M_1. He can do this because this is a 1D search, which we have just seen is feasible. Once at M_1, he determines the direction of steepest descent from that point and then goes off in that direction, performing a 1D search, until he again finds a minimum. He continues this process until he no longer finds himself getting significantly lower. It is also possible to reassess the direction of steepest descent after each step, rather than after reaching the minimum in the plane determined by the first step.

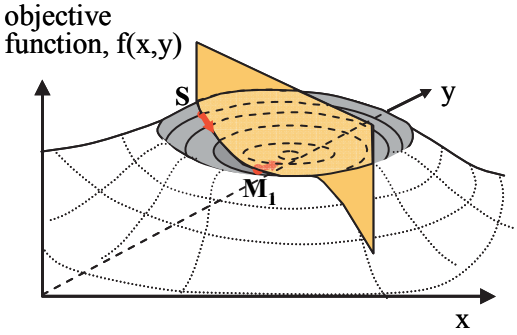


Figure 9.7. Using the method of steepest descent in a two-dimensional world – in this case a valley somewhat of the form of the crater of a volcano (see text)

One problem with the method of steepest descent is that it can be very inefficient. For example, if the searcher finds himself in a long narrow valley, he may find himself taking very many small steps as he continually crosses the valley from side to side, while inching toward the bottom. The conjugate gradient method largely solves this problem. It differs mainly in its way of deciding on the next direction to go after each step. Its algorithm to do this is based on finding good directions in which to go that do not "interfere" with the direction

along which the previous minimization occurred. It largely avoids, therefore, the criss-crossing behavior that the method of steepest descent exhibits in long narrow valleys. This scheme is not very intuitive, but it has the great advantage that it works. It is widely used for optimization problems in radiotherapy.

Global vs. local extremum

There is a fly in the ointment. The landscape in the previous figures showed only one deep valley. The algorithm described could reliably find it. However, Figure 9.8 shows a more unfortunate scene. Two minima are apparent – and there could be large numbers of them if the score function has any complexity. The value of the score function at the point labeled L is a minimum within a restricted region. Such a low point is termed a *local minimum*. But the value at L is not the lowest value everywhere. That value, in this example, is at G. It is termed the *global minimum*. Obviously, the 1D search method described above would have a very good chance of finding L and then stopping, thus missing finding the global minimum.

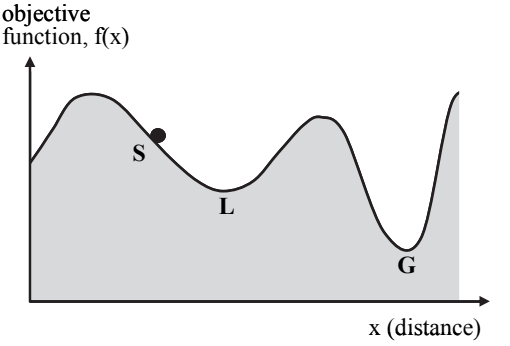


Figure 9.8. More than one minimum! Can one find the global minimum? (see text)

Iterative Optimization Using Mean-square Dose Deviations for the Score

There is an important example of iterative optimization (Bortfeld *et al*. 1990) in which the score measures how well the dose distribution conforms to a desired dose distribution. The measure of dose conformity is given by an equation of the form:

$$\text{score} = \sum_{i \in \text{target}} \text{weight}_i \cdot (D_i - D_i^{\text{prescribed}})^2 + \sum_{i \in \text{normal tissues}} H(D_i - D_i^{\text{limit}}) \cdot \text{weight}_i \cdot (D_i - D_i^{\text{limit}})^2$$

In this equation the two sums are, respectively, taken over all voxels in the planning target volume and in normal tissues, and the subscript "i" refers to the i'th voxel. For voxels in normal tissues, D_i^{limit} is

chosen to be a clinically acceptable value. D_i^{limit} is an upper dose constraint which is enforced by the multiplicative function, H.[7] The weighting factors, $weight_i$, are importance factors. This formulation has an operational advantage in that the computation of the derivatives of the score function, which are needed to determine the direction of descent, is particularly easy and fast. It also has the very attractive feature that the score function has no local minima to get trapped in. This is perhaps the most widely used score function at this time.

At first glance, this might seem to be the same as the dose distribution goal of the analytic optimization process – but, it is not. There is a very fundamental difference. In analytic optimization, the algorithm tries to achieve a defined dose distribution and when, as always, it fails, it achieves a compromise through a mathematical modification which is buried in the algorithm. Iterative optimization uses the dose discrepancies in all voxels as the measure of how close it is to its goal but, on the other hand, achieves its compromises through the user's instructions as exemplified in the importance factors. By adjusting them appropriately, the score can be made more sensitive, or less, to discrepancies between the actual and the desired dose values in particular regions. That is, there is a place in the process for introducing clinical tradeoffs.

Simulated Annealing

The method of simulated annealing is a quite different search procedure that has been developed in recent years and shows great promise for solving many otherwise intractable optimization problems. It has the great appeal that it can, in principle, find the global minimum. Stated more rigorously, given an infinitely long time and some other conditions, a simulated annealing search is guaranteed to find the global minimum. Unfortunately, we who are in the business of treating cancer generally cannot wait that long, so this guarantee is less impressive than it at first appears.

The process gets its name and rationalization from observations of the way materials anneal as they are cooled. However, this analogy is not

[7] H(x) is the so-called "Heviside step function" whose value is 1 for positive values of x and zero for negative values of x. Thus, it only introduces a penalty into the score function if the dose in a voxel exceeds the desired dose limit.

needed to formulate or understand the method. Simulated annealing differs from gradient search techniques in that it does not try to find the fastest way down the slope. Rather, it involves making random guesses as to where the minimum is.

This process is shown schematically in Figure 9.9. Imagine that, as usual, we start at point S. We make a random guess at where a "better" point might be, drawing our step size randomly from a distribution characterized by a "throw parameter". If the guess yields a point, such as P_1, at which the value of the score function is less than it is at S (throw "1" in Figure 9.9), then we accept it, move to P_1, and start the process over again. On the other hand, the guess may yield a point, such as P_2 or P_3, at which the value of the score function is more than it is at S (throws "2" and "3" in Figure 9.9). In most optimization schemes, such points, being worse, would be immediately rejected. In simulated annealing, paradoxically, one sometimes accepts such a point. With some probability whose magnitude is randomly drawn from a distribution characterized by a "cooling parameter", one may elect to accept the worse solution – although, of course, P_1 won't be forgotten; it could yet be our candidate for the minimum. That is, we have a chance of moving to P_2 – which won't be all that great – or, much more promisingly, to P_3 – a point from which we obviously have a much better chance at ending up at G, where we'd like to be.

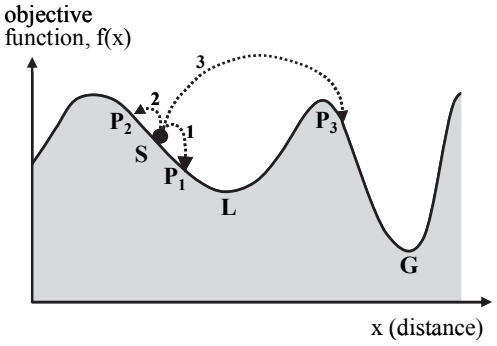


Figure 9.9. Simulated annealing: random guesses are made as to where the minimum is – and sometimes "uphill" (i.e. apparently worse) guesses are accepted (see text).

In order to converge, as the process proceeds, one must both reduce the average distance over which throws are made, thus reducing the incidence of wild throws, and must reduce the probability of accepting uphill throws. The parameters that control these must not only be given starting values; they also need to have a defined schedule by which they are reduced.

There are variants of simulated annealing, such as the so-called fast simulated annealing method. This differs from simulated annealing as just described in that: (1) the distribution from which random throws are drawn is not the Gaussian distribution but the so-called Cauchy distribution; and (2) the cooling schedule is much faster.

Pareto optimization

A recent approach to planning is the use of so-called Pareto optimization (Bortfeld, 2003). Pareto optimization involves modification, not of the usual input variables such as gantry angle, beam weight and so on, nor of importance factors, but, rather, of measures of *output* quantities such as EUD or TCP or NTCP which quantify the effect of the plan on the VOI's of interest; let us call these measures the output variables. In this approach, a vast number of plans are generated by computer. Plans for which improvement of one of the output variables will inevitably result in the worsening of at least one other output variable are said to lie on the Pareto front. The planner views an interactive display which shows the values of all the output variables for the currently selected plan. He or she can then increase or decrease the value of any one output variable – thus moving to a neighboring plan on the Pareto front – and view, interactively, the consequences for all other output variables. For example, one might reduce the NTCP for the spinal cord and see what impact that would have on the TCP and the NTCPs of other critical structures. In this manner, the user can choose a plan that represents his or her best judgment of the most acceptable plan amongst those which lie on the Pareto front.

This approach in no way evades the problem of subjectivity in the comparison of plans. However, it has two attractive characteristics: it constrains the user to view only a productive subset of possible plans; and it allows the user to make adjustments in the space of clinically meaningful variables. These adjustments are the tradeoffs, discussed briefly in Chapter 6 and below, that are at the heart of the optimization process. The great contribution of Pareto optimization is that it makes the tradeoff process explicit and, hence, more transparent to the planner.

Some issues in mathematical optimization

I want now to draw your attention to a number of issues that arise in the context of optimization. Optimization appears on the face of it to be an automated process. However, in fact, it often needs considerrable hand-tuning, specific to the application, to make it work well.

Starting Values

In principle, an important insight into the search procedure can be gained by observing the behavior of the algorithm when one *changes the starting point* and restarts the search.[8] An ideal search process would arrive at the same result, or at least a clinically comparable result, regardless of where one started it. If this happens, one can be relatively content. If any reiteration of the search yields a substantial improvement over the best previous result, then it has "hopped over" an intermediate hill and dropped into a lower valley. There is some chance that this new valley contains the global minimum. When using searching algorithms, several searches, each using a different starting value, should be attempted and their results compared. However, this is rarely done in practice, both because of time constraints and because the solutions found using the initial set of starting values are often satisfactory, even if not optimal.

Scale

There is a problem of "scale." Many treatment variables, such as distance, angle, intensity etc., have different units and quite different ranges - e.g., 0 to 20cm for a collimator setting, 0-360° for a gantry angle, 0 to 2 Gy for a pencil beam weighting factor, and so forth. The size scale for a step has to be established independently for each direction. For example, one might pick step sizes of 3 mm for collimator changes, 5^{o} for gantry angle setting, etc. The problem of scale is also evident in establishing the extent of the search space. Within what spread of values of each variable will one pick a starting point? If that spread is too small, and step sizes are too small, it might not be possible to reach an extremum in a reasonable period of

[8] The user often is unaware of what starting values are used in the optimization program being used. In the early days of iterative optimization, the analytic inverse optimization result was sometimes used for the starting values.

time.[9] There is no universal answer to these questions. They have to be answered in the context of the problem being solved. When dealing with an optimization scheme for radiotherapy, one hopes that these decisions can be made once and for all and be embedded in the algorithm so that they do not need to be revisited, but this may not always be the case.

Search Parameters

Even if, for our problem, scale issues do not need to be continually readdressed, there are parameters of the search process which may well need to be adjusted to make it work well for a particular case. For example, in the simulated annealing algorithm the behavior of the search is quite strongly affected by the initial size of the cooling and throw parameters and by the schedule for their modification as the search proceeds. These may need to be adjusted if the search does not appear to be converging to a solution.

When is an iterative search over? I have made several references to ending a search when the score is no longer significantly improving. But, what does the term significant mean in this context? Is a change of, say, 0.1 in the score large or small? Of course, one cannot answer this question without an understanding of what the score represents; there will be quite different answers for different score functions, or for the same score function with different importance weighting factors. This question requires the user's expert understanding of the nature of the score function, largely gained from making numerous similar searches in the past.

When many different starting points are tried, or different score functions are used, it is a common experience that the solutions are

---

[9] The importance of understanding the scale of step sizes used in a search can be seen by analogy with the following scenario. Imagine that Figure 9.6 represents a countryside landscape, and that the distance between the starting point, S, and the location of the low-point, M, is some hundreds of meters. If a hiker starts walking down-hill from S, taking normal meter-long strides, he or she will have an excellent chance of locating M. Imagine, on the other hand, an ant starting out from the same point, but taking millimeter-sized steps. The ant is very likely soon to find him or herself in a very small indentation in the earth and, since he or she measures slopes over very small dimensions, may conclude that the minimum has been found. Yes, it's a minimum, but not the global one. And even if we assume the surface is without small indentations, the hapless ant will take an extraordinarily long time to get to the bottom.

different from one another in the sense that the sets of treatment variables that constitute the solutions have different values. However, it is very often the case that, while the variables have different values, the scores are very similar, and the corresponding dose–distributions may even look much the same. This similarity may mean, in essence, that the valley in the region of the minimum is quite flat. Two solutions may have a very similar height in the search space, but be quite far apart. This should not be a concern; it means that the solutions are clinically almost as good as one another.

Re-optimization and Tradeoffs

The planner often finds that the results of a search are unsatisfactory, for one reason or another, and has to repeat computer-automated optimization several times. It is an odd sort of optimization which, when used the first time, is likely to appear to a planner to be so sub-optimal that the "optimization" has to be repeated.

And, what is changed when repeating the process? Well, as I previously warned, the planner's folded hands as seen in – or, rather, inferred from – Figure 9.5, are a bit illusory. As already emphasized, there are parameters of the optimization that must be supplied by the planner and which he, therefore, has the power to modify.[10] Among these are: the values of the dose constraints; the relative weightings of elements of the score function, when there are more than one; the dose goals; and the importance weighting factors. These parameters determine the tradeoffs which the searching process has, in effect, made among the various objectives. Tradeoffs are at the core of the planning process, as was briefly discussed in Chapter 6. The issue of tradeoffs in IMRT has been addressed by Hunt (2002). Planners often find that some parameters require adjustment in order to produce an acceptable plan – that is, one whose dose-distribution is deemed desirable.

What should one blame for the phenomenon in which clinically acceptable results are not achieved on the first attempt? Is it due to bad score functions, or bad search procedures? Most search procedures are, *per se*, fairly neutral with respect to the biology of the problem; the biology is in the score function. I believe therefore that the problem, which is one of clinical acceptability, usually comes

[10] Not to mention all the non-optimized parameters he or she is responsible for setting.

from having an inadequate score function. This only emphasizes the point that I made before, namely that it is much harder to develop a good score function than it is to find out how to search for an extremum of it. We will know that our score functions are good when we find that re-optimization is no longer necessary.

Biology Buried in the Mathematics

The mathematics involved in optimization is relatively modest in a technical sense but can, nevertheless, be quite complex and details may be hidden from a user's eyes, for many reasons. My concerns are:

- Some algorithms, while wrapped in the guise of fulfilling a clinically motivated prescription, are in fact mathematical formulae *with no biological basis or content* whatsoever.
- The modest mathematical complexity of some models may obscure both their physical and biological content or lack thereof.
- Worst of all, the impenetrability of the algorithms may leave some users no choice but to uncritically accept their conclusions.

When using automated optimization, it is the user's responsibility to his or her patient to understand and be satisfied with what lies under the covers, so to speak. *Caveat emptor*.

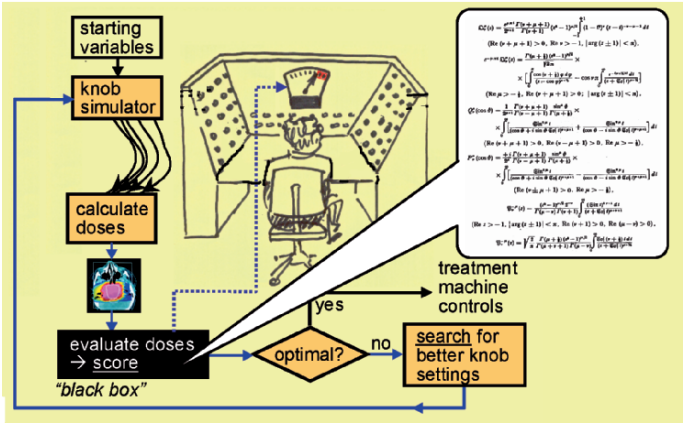


Figure 9.10. Hush. Optimization in progress.

OPTIMIZATION?

Voting the best piece of music

Every year in Boston – and, I am sure, in towns all around the world – a local classical radio station asks its listeners to vote for the piece of classical music that they consider to be the best. They get thousands of responses from listeners who certainly know that the question is nonsense. There are so many dimensions to a piece of music, so many facets, and so many different occasions for which different kinds of music might be preferred, that picking the "best" is, at most, an amusing exercise. Optimization in this setting is simply neither feasible nor useful.

Is treatment planning very different? There are, indeed, many dimensions to the problem, and many of them are expressed in terms that are unrelated to the way others are expressed. How shall one combine a measure of the distribution of dose within the target volume with the V_{20Gy} in a normal tissue? These are different facets of the treatment, expressed in different physical units. This conundrum makes the job of summarizing all these facets into a single score, so that one plan can be ranked as the best, quite daunting. There is one big difference, however. The patient *must* be treated; one of the possible treatment plans *must* be selected; and that plan will be selected because, in the planner's judgment, it is the best that can be achieved in practice. A music director can pick this or that piece of music, as whim directs, and there will generally be only mild consequences if he misjudges his audience. A treatment planner has no such freedom. This means that the exigencies of medical care force optimization upon us, like it or not, whether or not the optimization scheme is "optimal."

The meaning of the term optimization

The term "optimization" has at least two rather different meanings and this causes considerable confusion. In its *vernacular sense* – that is, the meaning intended in everyday speech – it is the process of finding the best possible solution to a particular problem. In the *mathematical sense*, it is the process of finding values of the independent variables that lead to an extremum of a score function.

The reason that this distinction needs to be appreciated is that an optimized plan in the mathematical sense may not be optimal in the vernacular sense.

SUMMARY

I have presented numerous arguments in this Chapter to suggest that optimization schemes, as they exist today, are sub-optimal. To summarize the arguments:

- the score functions address only a subset of the treatment variables whose values must be defined for therapy, the others being selected manually by the treatment planner;
- the search process may not find the global extremum of the score function;
- we anyway don't know enough biology to definitively answer the question of which of perhaps thousands of plans is the best;
- the scoring scheme, that is the score function, may not reflect the planner's judgment very well, if at all;
- the problem is multi-faceted and one is hard put to know how to combine the different aspects into one score.

For all these reasons, I believe that *optimization in the vernacular sense is either unachievable, or will not be achieved in our lifetimes*. I have led you through the descriptions in this chapter, and presented you with my long list of issues, for one particular purpose. Namely, to instill in you, as a user of optimization programs, the need for considerable caution.

In treatment planning, optimization is a misnomer that can give rise to a false sense of complacency. What I believe *is* possible is to improve a plan through adjustment of some of its treatment variables from their starting values, using some of the techniques described above. For this reason I prefer the term *plan refinement* to plan optimization.

In the end, and despite all my qualifications, IMRT has led to the design and delivery of much better treatments than were possible without it. One can anticipate that techniques to further improve automated plan refinement will receive a great deal of attention in the coming years and I am sure that those efforts will result in even better treatments being given to our patients.

RADIATION ONCOLOGY

RADIATION ONCOLOGY: A PHYSICIST'S-EYE VIEW

Michael Goitein

Springer

Michael Goitein
Harvard Medical School
Boston MA, USA
and
Ankerstrasse 1
5210 Windisch
Switzerland

Library of Congress Control Number: 2007932210

ISBN 978-0-387-72644-1 e-ISBN 978-0-387-72645-8

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

for Gudrun

PREFACE

This book describes how radiation is used in the treatment of cancer. It is written from a physicist's perspective, describing the physical basis for radiation therapy, and does not address the medical rationale or clinical aspects of such treatments. Although the physics of radiation therapy is a technical subject, I have used, to the extent possible, non-technical language. My intention is to give my readers an overview of the broad issues and to whet their appetite for more detailed information, such as is available in textbooks.[1]

I hope that this book will be interesting and helpful to radiation oncologists, to physicists, and to those who are curious about, but not yet engaged in, the field. I hope, too, that experienced practitioners can get something out of this book. However, it is directed primarily toward those just entering the field, and to those contemplating entering it – either from the physics or the medical side. I have been asked, "Will it help residents pass their board examinations?" I can make no promises; it is certainly not a cook-book of answers. But I think it could help.

I have avoided formulae and quantitation so far as has been possible. I think that there is a schism between the descriptive methods that are useful in physics and those that are needed in medicine or biology. Physics is, to a large extent, a highly successful effort to explain physical phenomena through mathematical formulae. It is quite astonishing, for example, how much that happens around us is described by Maxwell's four relatively simple equations. Such successes may suggest to us that the formulae are a fundamental basis of the reality they describe – not just phenomenological approximations to it. It seems to me that physicists are so beguiled by the success of formulae

[1] There are many excellent textbooks dealing with both medical physics and radiation oncology. Reading this book is in no way a substitute for studying these. An excellent medical physics text is Johns and Cunningham (1983). Unfortunately, this book, which went through four very successful editions, is no longer being updated. As a result, it no longer covers the very latest developments in the field. There are a number of more up-to-date textbooks, such as that by Khan (2003).

in explaining much of the physical world that they are tempted to think that the same approach will work in the world of medicine and biology. But, I think that an understanding of many important medical and biological matters cannot be based on mathematical relationships, so I have avoided them where possible.

I have generally shown schematic, rather than quantitatively accurate, figures, and have stated only approximate values of various quantities of interest. I have tried to appropriately qualify my figures and statements within the text, but I would like to caution readers at the outset that the data presented here should not be used as a basis for the treatment of patients. Treatments must be based on measured, or at least confirmed, data appropriate to the local environment, which are interpreted by qualified experts.

I have concentrated my focus on radiation therapy using external beams of high energy X-rays and of protons. With regret, I have had to neglect important techniques such as brachytherapy (the use of radioactive materials either implanted or inserted into the patient) and electron beam therapy, and I have not been able to address the several specialized forms of external beam x-ray therapy such as radio-surgery, gamma-knife therapy, robotic therapy and tomotherapy. In a very few instances, I have discussed matters which are not yet part of mainstream practice. This is the case, for example, in my discussions of: the calculation and display of the uncertainty bounds of dose distributions; the use of Monte Carlo techniques to calculate dose distributions; and the implementation of pencil beam scanning to deliver intensity-modulated proton therapy. I trust that these will soon be routine.

Lastly, I have not done justice to the huge literature of the subjects I have covered. I have pointed to some few articles which seem to me to be of interest, but I have omitted many others of equal or greater value. And, I confess, I have tended to cite my own publications disproportionately since many of the issues I address in this book have been the focus of my own work and writings.

My wife, herself a radiation oncologist, was trained by a demanding and intellectually endowed man who asked a lot from his trainees and staff. She tells me that, knowing that everyone makes errors from time to time, he told them that he could accept any mistake they made so long as the one who had made the error knew *why* he or she had

done what they had done. I hope this book will arm my readers to better know *why* they can or should do, or not do, certain things. Unfortunately, in establishing safe and reliable procedures for patient treatments, medical physics has tended to suffer, in my opinion, from a certain cook-book attitude. But, for me, "Because we've always done it that way" is simply not an acceptable answer to the question "Why?"

My complimentary intention is to encourage the asking of the question "Why not?" So often one hears a nascent idea being unthinkingly dismissed as impractical, unreasonable, or impossible. This usually happens when someone on one side of the physics/medicine partnership proposes something novel to someone on the other side. My wish is to give people on both sides of the divide a sufficient understanding of the knowledge and methodology of the other side that they will not be afraid to ask "Why not?" when their next brainwave is summarily rejected. One should not give up on an idea until its critics can convincingly explain why it cannot or should not be done. It is partly my goal to encourage your questioning of everything in your discipline – and, not least, of my own words.

Almost all of us have, or will have, direct personal knowledge of cancer. In the United States, roughly two out of five people will, on average, get cancer during their lifetime. This means that there is, on average, a 96% chance that, of the eight people nearest to you amongst your family and friends, at least one will develop a cancer during his or her lifetime. So, cancer is important to all of us. And, radiation therapy is important to cancer – approximately half of all cancer patients will receive radiation as at least part of their therapy. It is therefore my hope that you may be motivated to look further into the fascinating field of radiation therapy, if you are not already in it, and to have a better perspective on it, if you have already committed yourself to it.

MICHAEL GOITEIN

Windisch, Switzerland
June 2007

michael@goitein.ch

CONTENTS

10. PROTON THERAPY IN WATER


Robert Wilson

The ideal radiation with which to treat cancer is one that delivers a defined dose distribution within the target volume (generally a uniform distribution, but possibly a non-uniform one) and none outside it. This is unachievable. The next best thing would be a radiation

that delivers most of the dose within the target volume and relatively little outside it. Protons come much closer than do external beam photons to accomplishing this desirable goal, as is immediately evident on comparing the broad beam dose distributions of protons and high energy photons shown in Figure 10.1. This was first realized by Robert Wilson in 1946 and, prompted by his seminal article (Wilson, 1946), protons and other light ions have been evaluated at a number of centers throughout the world over the past four decades for their promise of providing superior therapeutic radiation beams. *

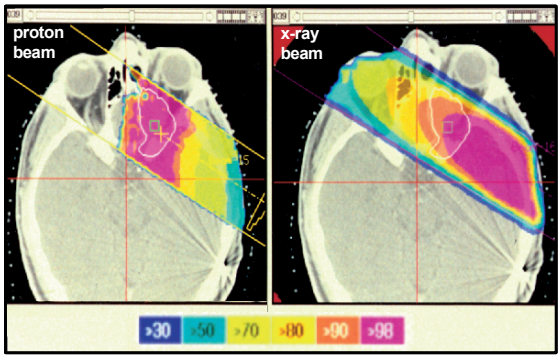


Figure 10.1. Side-by-side comparison of (left) a proton, and (right) a photon posterior-oblique beam.

To what is the dramatic difference, seen in Figure 10.1, between beams of protons and photons due? Well, protons, being charged particles, interact with matter very differently than do photons, whose interactions were discussed in Chapter 4. Their different modes of interaction result in quite different dose distributions. Now, my alert readers will already have recognized that the same is true of electrons, which are also charged particles, and whose interactions with matter were also discussed in Chapter 4. Indeed, the types of interactions that protons experience are quite similar to the interactions of electrons. However, protons are some 1836 times heavier than electrons and this has the consequence that proton dose distributions are quite different in practice.

* Some of the material in this chapter is adapted, with permission, from the article "Treating Cancer with Protons" which appeared in the September 2002 issue of Physics Today (pp 45–50) by Goitein M, Lomax AJ, and Pedroni ES. A good source of information concerning proton beam therapy can be found in ICRU78 (2007), from which portions of this chapter have been taken with the permission of the Oxford University Press.

THE PHYSICAL PROPERTIES OF PROTONS

When protons of a given energy pass through matter, they are subject to three main phenomena: Coulomb interactions with atomic electrons, Coulomb interactions with atomic nuclei, and nuclear interactions with atomic nuclei.

Coulomb interactions of protons with atomic electrons

Protons gradually lose energy, and hence deposit dose, as they penetrate matter. This energy loss is mainly due to Coulomb interactions of the protons with the orbiting electrons of atoms. The opposite charges of the protons and electrons cause the protons to attract the electrons and "suck" some of them out of the atoms. This results in ionization of atoms and, even more importantly, setting loose electrons that go on to ionize further atoms in the neighborhood of the initial ionization, just as was described in chapter 4. This process is shown schematically in Figure 10.2, and should be compared to Figure 4.6 of Chapter 4. On average, the protons lose relatively little energy in individual ionizations and are very little deflected; they suffer some 100,000s of interactions per centimeter of material before eventually losing all their energy and coming to rest.

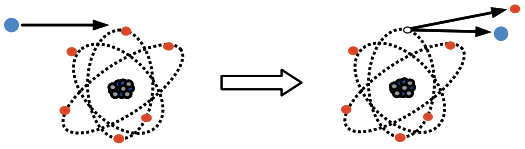


Figure 10.2. Coulomb interaction of a proton (*blue*) with an atomic electron (*red*). The particle sizes are schematic only; they are totally out of proportion in the figure. Indeed you could not see any of the particles if the figure was drawn to scale.

A monoenergetic proton beam will penetrate matter of a given density up to a well-defined depth, which is determined by the beam energy. The fact that the depth of penetration is related to the proton energy in a one-to-one manner is the key to the practical use of protons in radiation therapy, for it allows the penetration of the beam within the patient to be controlled, at the sub-millimeter level if necessary, by simply controlling the energy of the protons incident upon the patient.

Coulomb interactions of protons with atomic nuclei

Because protons are so much heavier than electrons they are hardly deflected at all by Coulomb interactions with atomic electrons – as just described. However, they also experience a repulsive force when they pass close to a positively charged *nucleus* of an atom (see Figure

10.3). In Coulomb interactions with *nuclei* (as opposed to with electrons), because the nuclei are so much heavier than electrons, they can deflect protons through larger, although still small, angles. A proton suffers very many such interactions with atomic nuclei as it passes through matter and the deflections caused by all these interactions add up statistically, resulting in a net angular and radial deviation. This important phenomenon is termed "multiple Coulomb scattering."

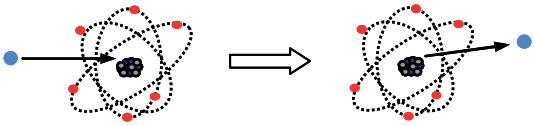


Figure 10.3. Schematic representation of Coulomb scattering of a proton (*blue*) by an atomic nucleus.

Bremsstrahlung

Just as in the case of electrons, protons when passing in the field of an atomic nucleus suffer a lateral acceleration that, since they are charged, results in the emission of a spectrum of photons. The difference is that the likelihood of bremsstrahlung is roughly proportional to the inverse of the square of the particle mass. As a consequence, proton bremsstrahlung is more than a million times less intense than electron bremsstrahlung and is not of any clinical significance in proton beam therapy.

Nuclear interactions of protons with atomic nuclei

In addition to Coulomb interactions with atomic nuclei, protons suffer nuclear interactions with them, via the so-called "strong nucleon-nucleon force." There are, in general, two important types of nuclear interaction:

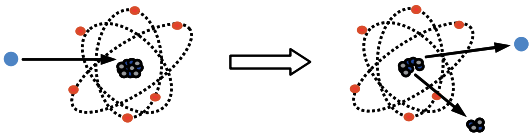


Figure 10.4. Schematic representation of a non-elastic nuclear collision of a proton (*blue*) with an atomic nucleus leading, in this example, to break-up of the nucleus with emission of an alpha-particle.

- elastic collisions with a nucleus in which the nucleus is left intact but the proton loses a significant fraction of its energy and is usually deflected by several degrees (e.g., $p + {}^{16}O \rightarrow p + {}^{16}O$).
- non-elastic collisions with a nucleus in which the nucleus is broken apart and the incoming proton loses a significant fraction of its energy and is usually deflected by several degrees (e.g., $p + {}^{16}O \rightarrow p + {}^{15}N + p$).

In non-elastic collisions, the nucleus may disintegrate in a number of ways, but generally a relatively light fragment is knocked out with considerable speed leaving behind a heavy fragment which stays close to where the interaction took place and is heavily ionizing, as illustrated in Figure 10.4. The relative energy carried away by the fragments that are produced is given in Table 10.1.

Table 10.1. Fractional energy loss taken up by various particles when 150 MeV protons strike a ^{16}O nucleus. Data taken from Selzer (1993).

particle	*fraction of energy (%)*
protons	57
neutrons	20
alpha particles	2.9
deuterons	1.6
tritium	0.2
helium-3	0.2
other charged recoil fragments	1.6

THE DEPTH–DOSE DISTRIBUTION OF A BROAD PROTON BEAM

I now want to address the dose characteristics of a broad beam of protons with a uniform lateral intensity distribution. Such a beam is produced by passive scattering (see below), but it can equally well be produced by scanning by keeping the weights of all pencil beams of a given energy the same (see below).

The Bragg peak

The dose deposited by protons rises sharply near the end of their range, giving rise to the so-called Bragg peak, named after Sir William Henry Bragg (who should not to be confused with his son, Sir William Lawrence Bragg, also a physicist.) An example of a typical dose distribution of a near-monoenergetic proton beam is shown in Figure 10.5.

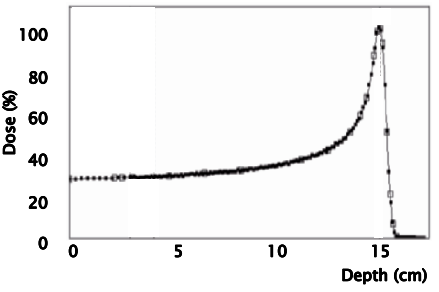


Figure 10.5. Depth dose distribution of a mono-energetic ~150 MeV proton beam in water, showing the characteristic Bragg peak. Figure courtesy of B. Gottschalk, HCL, USA.

The shape of this distribution arises from a number of cooperating effects, which I now describe.

Energy loss due to Coulomb interactions with atomic electrons

The principal genesis of the Bragg peak is the slow loss of energy that protons experience due to Coulomb interactions with atomic electrons which, you will recall, cause protons to slowly lose energy by transferring it to atomic electrons. However, they do not give up energy equally at all depths. At any given point within a stopping medium, a proton's linear rate of energy loss – its "linear energy transfer" (LET), or "stopping power" – which is measured in units of MeV per $g{\cdot}cm^{-2}$ – is given by the Bethe-Block formula.[1] It is approximately proportional to the inverse of the square of the proton's mean speed, v:

$$\frac{dE}{dx} \propto \frac{1}{v^2}\left(\frac{Z}{A}\right)z^2 \qquad (10.1)$$

where Z and A are, respectively, the atomic and mass numbers of the target nucleus and z is the charge number of the projectile proton. The local energy deposition (i.e., the dose) thus rises sharply as protons slow down. This slowing down process, with its concomitant increase in dose with depth, is depicted schematically in Figure 10.6. Suppose the proton speed at a point such as A is v_A, then the dose it would deposit at A would be given by equation (10.1), substituting v_A for v. At a deeper point such

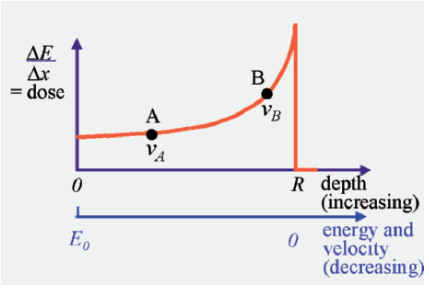


Figure 10.6. The contribution to the Bragg peak from Coulomb scattering of protons off atomic electrons (see text).

[1] In describing depths of penetration, the *areal density* is often employed, with units of $g{\cdot}cm^{-2}$. The areal density of a uniform medium is the product of path length and density – or, in inhomogeneous media, the integral of density over the path length. This integration removes the trivial (from a physics point of view) influence of density. Trivial because if one could, say, double the density of some medium and simultaneously halve its thickness, its effect on a proton beam would be virtually unchanged. For water, which has almost unit density, the areal density is numerically equivalent to the depth of penetration measured in centimeters.

as B, the proton will have slowed down, so that v_B will be smaller than v_A. Consequently, the dose at B, evaluated according to equation (10.1), will be larger than that at point A – and, indeed, considerably larger, the closer A is to the end of range, since the velocity is close to zero near the end of range. At the end of range, protons come to a stop and the dose therefore drops precipitously to zero, resulting in the highly asymmetric peak of Figure 10.6.

Range straggling and energy spread of the incident beam

The ionization peak of Figure 10.6 is blurred out by two effects. First, there are statistical fluctuations in the ionization processes. These cause what is called "range straggling" – a smearing out of the depth of penetration of stopping protons, typically by about 1% of their range. Second, one virtually never has a monoenergetic beam of protons; there is always some energy spread due to details of the protons' production; this energy spread is also typically of the order of 1%. These two effects combine in a near-Gaussian spread function and smear out the near-infinitely sharp ionization peak of Figure 10.6, resulting in the broader, more rounded and more symmetric peak seen in Figure 10.5.

Nuclear interactions

Finally, we need to take the nuclear interactions of the protons into account. They occur at a rate of ~1% per $g \cdot cm^{-2}$ up until the last few millimeters of their end of range (there is a threshold for nuclear interactions of around 20 MeV). Nuclear interactions have a number of consequences. They:

- gradually reduce the number of primary protons in the beam; a 160 MeV beam loses about 20% of its protons in this manner by the time the end of range is reached;
- produce a halo of scattered primary protons and knocked-out secondary protons that travel long distances, though not past the primary protons' end of range, and add a "tail" to the lateral dose profile of a beam;
- create heavily ionizing fragments with a very high stopping power that deposit dose very close to the point of interaction and increase the relative biological effectiveness (RBE) in the neighborhood of the interaction (see Chapter 11); and
- create a halo of neutrons that largely escape the patient without further interaction, but are responsible for a small contribution to the dose inside and outside the primary radiation field.

The relationship between the depth–fluence distribution (i.e., the number of protons as a function of depth) and the depth–dose distribution of a near monoenergetic proton beam is schematically illustrated in Figure 10.7. A useful thing to know is that the 80% dose on the falling edge of the Bragg peak is very closely at the same depth as the 50% fluence of the falling edge of the fluence distribution.

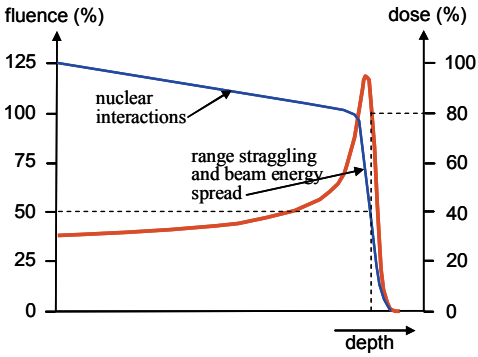


Figure 10.7. Schematic graph of the depth–fluence (*blue*) and depth–dose (*red*) distributions of a typical proton beam at therapeutic energies.

The near-flatness of the depth–dose distribution of protons in the entrance plateau region (e.g., as seen in Figures 10.5 and 10.7) is a purely fortuitous cancellation between the rising energy deposition seen in Figure 10.6 and the diminishing number of primary protons due to nuclear interactions as seen in Figure 10.7.

Bragg peak dependence on energy and energy spread of the beam

The *range* of a proton beam is best defined as the depth of penetration from the front surface of the stopping medium to the distal 80% point on the Bragg peak (relative to 100% at the top of the Bragg peak). The depth at which the Bragg peak occurs depends on the initial energy of the protons; the greater the energy, the greater the range. The penetration of some selected proton beam energies is shown in Table 10.2 and in Figure 10.8.

Table 10.2 Range of mono-energetic protons in water.

energy (MeV)	*range in water (cm)*
70	4.0
100	7.6
150	15.5
200	25.6
250	37.4

In addition to having shorter ranges, lower energy proton beams typically have narrower Bragg peaks, as can be seen in Figure 10.8. Range straggling of the protons and energy spread in the beam, as just mentioned, spread out the Bragg peak by some percentage of its range, typically about 1.5%, more or less independent of the proton energy. The Bragg peak is narrower at lower energies because the near-constancy of the width of the Bragg peak relative to its range translates into a smaller *absolute* broadening of the Bragg peak at shorter ranges (i.e., at lower energies).

Lower energy beams have a higher peak-to-plateau dose ratio (the ratio of the dose at the peak of the Bragg peak to that at near-zero depth), as seen in Figure 10.8. This phenomenon is driven by the just-discussed narrower widths of the Bragg peaks of lower energy protons. No matter what the incident proton energy may be, just about the same amount of energy is deposited in, say, the last couple of $g \cdot cm^{-2}$ of a proton's path in a medium. Thus, a narrower peak has to be higher in order for the total energy in the peak to be constant. In consequence, lower energy beams, since they have narrower Bragg peaks, have a higher peak-to-plateau dose ratio than higher energy beams.

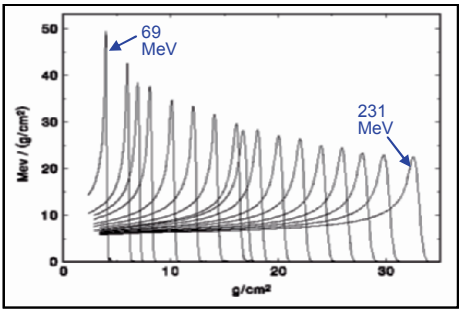


Figure 10.8. A series of Bragg peaks of beams with energies of between 69 and 231 MeV. Figure courtesy of B. Gottschalk, HCL, USA (Gottschalk, 2004).

At an energy of around 150 MeV, the value of the peak-to-plateau ratio is about 3:1 in practice. At that energy and in water-equivalent material, the width of the peak is typically about 6 mm at the 80% dose level, and the distal falloff of dose from 80% to 20% – which is the typical descriptor of "penumbra" – is about 4 mm. The distal falloff becomes less steep at higher energies, following the broadening of the Bragg peak – and, conversely, becomes steeper as the energy is reduced. Thus, the "distal penumbra" of a 200 MeV beam whose range in water is 25.6 cm, as shown in Table 10.1, is a relatively broad 7 mm, while that of a 70 MeV beam whose range in water is 4.0 cm is close to 1 mm.

If the penetration of protons is selected by placing material in the beam just upstream of the patient (termed a degrader), rather than by changing the beam energy, then the preceding comments about the Bragg peak width and the peak-to-plateau ratio no longer hold. The depth–dose distribution of the degraded beam in water is essentially the same as that of the un-degraded beam, but shifted towards smaller depths by an amount equal to the water-equivalent thickness of the degrader in the beam.

The spread-out Bragg peak (SOBP)

It was the dose distribution of a monoenergetic proton beam such as that shown in Figure 10.5 that so attracted Robert Wilson's attention (Wilson, 1946) and led him to suggest that a beam of protons would deposit almost all of its energy within a deep-seated tumor, none beyond it, and very little proximal to it. However, as we have just seen, the Bragg peak is very narrow. Few tumors are that small in extent. Most tumors used to be described, before the advent of more quantitative imaging methods, variously as plum-sized, orange-sized, and so forth. That is, they are likely to extend at least many centimeters in depth and sometimes more than 10 cm. To treat such tumors, the extent of the high dose region needs to be much greater in depth than is provided by a single Bragg peak.

As Wilson observed, an extension in depth can be achieved by delivering not just one, but many Bragg peaks, each with a successively slightly different range (i.e., energy). These peaks should not all be equally weighted. Rather, the more proximal a peak is, the less weight it should have. This is illustrated in Figure 10.9. The distal region of near-constant high dose is referred to as the *spread-out Bragg peak*, abbreviated as SOBP. Just how this non-uniformly weighted stacking of beams is accomplished is discussed below.

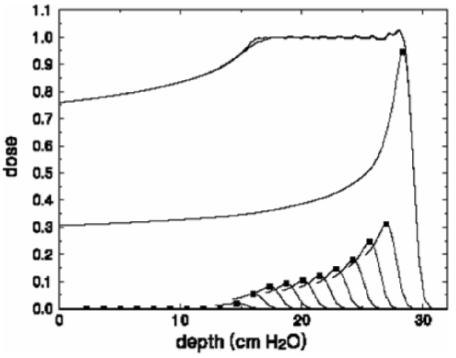


Figure 10.9. Illustration of how a spread-out Bragg peak (*top curve*) is made up of near-monoenergetic proton beams of successively lower energy and weight. Figure courtesy of B. Gottschalk, HCL, USA (Gottschalk, 2004).

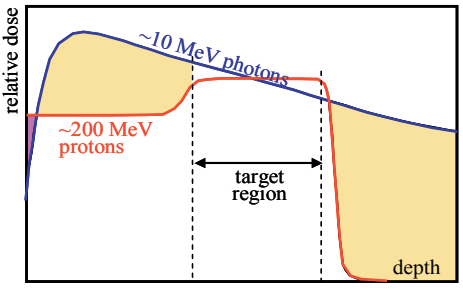


Figure 10.10. Schematic comparison of high energy photon and proton beam depth distributions. The golden area identifies unwanted dose delivered by photons, but not protons. The purple area at small depths identifies a small but sometimes important region in which protons lack the skin-sparing advantage provided by high energy photons.

Unfortunately, the SOBP, while still delivering virtually no dose beyond the high-dose region, delivers a substantial dose proximal to it. The entrance dose depends on the extent in depth of the SOBP and, to a lesser extent, on its maximum penetration; typically it can be 80% or even higher. Thus, in practice, one does not achieve the dream of having a beam that deposits significant doses only in the region of the tumor. Nevertheless, the dose distribution of a proton SOBP is much superior to that of a photon beam from a typical linear accelerator. Figure 10.10 illustrates the most important ways in which the two modalities differ.

There is one further effect that bears mentioning, which occurs equally in photon therapy, namely, the inverse square effect. Relative to an initially parallel beam, the flux of protons, and hence the dose, at a point a distance 'r' from the source (which may be either real, as in scattered beams, or virtual, as in scanned beams) will be reduced by a factor of $1/r^2$. The inverse-square effect depresses the dose at larger depths. This is compensated for by adjusting the weights of the upstream Bragg peaks, but this compensation unavoidably raises the entrance dose. The practical consequence of this is that the modest sparing of proximal tissues that a SOBP provides is further reduced, the shorter the source-to-patient distance is. For example, for a 250 MeV beam with a SOBP width of 10 cm, the ratio of the entrance dose to dose within the SOBP is: 64% at 30 m source-to-isocenter distance; 77% at 3 m; and 127% at 1 m. For this reason, one tries in proton beam therapy to keep the virtual source at least 2 m from the isocenter and preferably 3 m or more.

For scanned beams, if the scanning is achieved by a pair of dipole magnets offset from one another in distance along the beam axis, there will be two virtual sources at rather different distances from the patient. This somewhat complicates calculation of the inverse-square

effect, but its primary impact is to complicate such things as presenting a beam's-eye view image of the patient's anatomy and computing the tapering of beam trimmers, if any.

THE ELECTRON'S BRAGG PEAK

The depth–dose distributions of electrons in the therapeutic range of energies, which is generally from 6 to 25 MeV, appear to be smooth, decreasing monotonically towards the end of range, and peak-free. Why is it that, if the electron is simply a lighter cousin of the proton so far as its electromagnetic interactions are concerned, electrons don't exhibit a Bragg peak?

The answer is, I think, informative. The fact is that electrons do have a Bragg peak, but it gets blurred out to the point of vanishing. As you know from Chapter 4, electrons, like protons, gradually lose energy due to their Coulomb collisions with atomic electrons and, like protons, their rate of energy loss rises as the electron's energy decreases. This increase in dose with depth is the necessary condition for a Bragg peak to appear. Electrons, like protons, also experience multiple Coulomb scattering. The big difference is that, owing to their much lighter mass, electrons are scattered, and their path thereby altered, much more than protons. While protons scatter a few degrees and follow an only slightly un-straight path, electrons can be scattered through very large angles so that their path is dramatically modified, even to the extent that some electrons turn back on themselves.

If one followed an individual electron along its meandering track, and noted the dose deposited per unit path length, one would indeed observe a Bragg peak. However, for a beam comprised of very many electrons, it is their deposition of dose with depth in the medium, as opposed to path length, that can be measured and that is of therapeutic interest. Due to their large degree of multiple scattering, electrons crossing a plane normal to the beam and located at depth, are at a rather different points *along their paths* and so have quite a wide range of energies and, hence, of stopping powers. The dose one would measure at that plane, then, would be the average of the dose delivered by each of the electrons – and that averaging process blurs out the Bragg peaks to the point of invisibility.

Let us now return to protons, with no more parenthetical side trips.

THE DEPTH–DOSE DISTRIBUTION OF A SMALL DIAMETER BEAM

Below a diameter of about 15 mm, the depth–dose distribution of a small-diameter proton beam is substantially degraded as compared with that of a broad beam, as shown in Figure 10.11. How does this diminution of the Bragg peak come about?

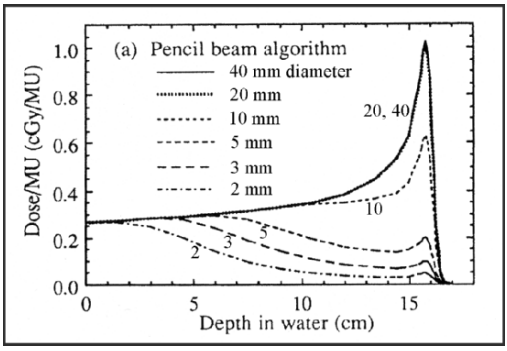


Figure 10.11. Depth–dose distributions of beams of varying diameter. Reproduced with permission from Hong *et al.* (1996).

To begin with, let us consider *pencil beams*. "Pencil beam" is a somewhat loose term, and it is useful to consider two types: (1) an *infinitesimal pencil beam* which is a beam whose size, angular divergence and energy spread at the point that the beam impinges upon the patient are infinitesimally small; and (2) a *finite pencil beam* for which the above parameters are not infinitesimal, but are nevertheless small compared, say, to the area of the field.

The near-disappearance of the Bragg peak of a small pencil beam is caused by multiple Coulomb scattering of the protons. If there were no such scattering, the depth dose distribution of a pencil beam would be no different than that of a broad beam. However, multiple Coulomb scattering causes protons to spread out laterally, and the deeper protons have penetrated, the more they have been scattered, and hence the more they are spread out. As a consequence, the energy deposited at the depth of the

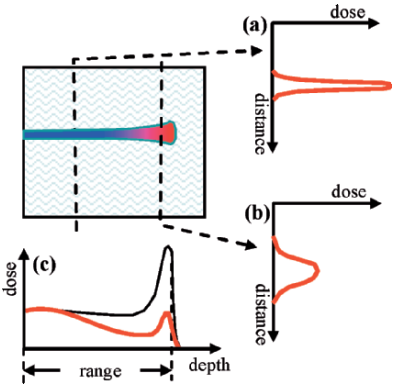


Figure 10.12. Single pencil beam, illustrating how the lateral dose distribution at large depths is broader but shallower than at small depths, resulting in a depression of the central axis dose at large depths.

Bragg peak is smeared out laterally much more than at shallow depths. Since protons are not "lost" but, rather, are spread out, the fluence at the end of range is of lower amplitude but of greater lateral extent than at shallower depths, as indicated in Figure 10.12c. For this reason, a small proton beam is not well-suited to the treatment of very small (e.g., a few mm diameter) but deep (e.g., many cm) target volumes.

A broad beam can be considered to be made up of a superposition of pencil beams, side by side. How is it that one can superpose pencil beams with miniscule Bragg peaks and end up with a broad beam with a large Bragg peak?

Figure 10.13 portrays a broad beam as being composed of a large number of pencil beams, set side-by-side. Since the pencil beams are very little spread *at small depths*, the point P will receive dose from probably only the one pencil beam pointed directly at it. On the other hand, since the pencil beams are considerably spread out *at large depths*, point Q near the end of range will receive dose not only from the pencil beam pointed directly at it, but from several adjacent pencil beams pointed at laterally adjacent points. The doses from all these beams add up to a much greater value than that which would be due to the directly-pointing pencil beam alone. Indeed, they add up to the broad-beam value.

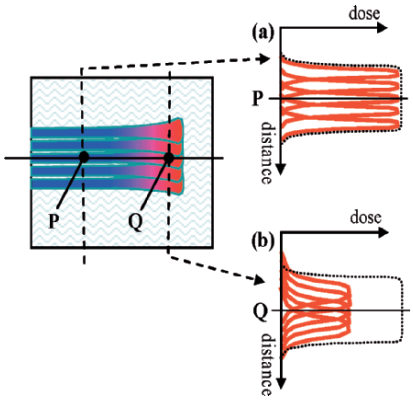


Figure 10.13. A broad beam made up of an array of parallel pencil beams. The dotted lines in the graphs on the right hand side are the sum of the doses from all pencil beams (see text).

Having said all this, of what use are pencil beams? It is twofold. First, the proton beams used for scanning, as discussed below, are finite pencil beams. An understanding of the properties of pencil beams is therefore essential for planning the delivery of a scanned beam. Second, the pencil beam is a useful theoretical concept for computing the dose delivered by a scattered broad beam within the patient, since it can be thought of as being composed of a number of pencil beams.

THE LATERAL DOSE DISTRIBUTION OF A PROTON BEAM

So far, we have mainly discussed the distribution in depth of the dose deposited by protons along the central axis of a beam. Now, let us see what happens at off-axis points.

Pencil beams

In the previous section, in discussing the near-disappearance of the Bragg peak of an infinitesimal pencil beam, I explained the phenomenon in terms of broadening of the distal end of the pencil beam caused by multiple Coulomb scattering of the protons. It is now time to look more closely at the causes of that broadening. There are two regions in which broadening occurs – namely, within the patient and in the material upstream of the patient. The four main effects that cause the broadening are as follows.

Multiple Coulomb scattering: near-Gaussian core

The details of multiple Coulomb scattering were worked out in around 1947 by Molière in a pair of comprehensive papers which have been more often quoted than read – see Gottschalk *et al*. (1993) for a discussion of Molière's theory. Multiple Coulomb scattering is the principal cause of the spreading out of an initially infinitesimal pencil beam. But multiple scattering comes in what can be taken as two separate parts.

The principal component is a nearly Gaussian distribution, both in the angle of deviation and in the consequent lateral spread of a pencil beam. Near the end of range, the standard deviation of the lateral distribution is approximately 2% of the range.[2] That is, a 150 MeV proton beam at its end of range (i.e., ~15 $g \cdot cm^{-2}$) will spread out sideways to form a near-Gaussian profile whose sigma is about 3mm and whose full-width at half maximum is about 7mm. As we have

[2] The following are some useful relationships. The full-width at half-maximum of a Gaussian distribution (with a standard deviation of σ) is 2.35σ; the 80-20% fall-off down one side of a Gaussian is 1.12σ; and the 80-20% fall-off of an error function (which is the shape that is generated when a set of equally weighted Gaussian distributions are summed up) is 1.68σ. This last is the number that characterizes the penumbra of a beam made from a sequence of equally spaced and weighted pencil beams whose lateral shape is a Gaussian.

already discussed, the extent of the lateral blurring is a function of where one is within the beam, being less at shallower depths than at the Bragg peak.

Multiple Coulomb scattering: long tail

However, that is not the whole story. The profile of an infinitesimal pencil beam of protons due to multiple Coulomb scattering is not precisely Gaussian in shape. There is a long tail that is due to large angle scattering in one or only a few collisions (Gottschalk *et al.*, 1993). This tail is of relatively low amplitude and can be approximated by a second, broader, Gaussian distribution for most purposes in proton beam therapy (Pedroni *et al.*, 2005).

Nuclear Interactions: protons

You will recall that both elastic and non-elastic nuclear collisions produce three classes of secondary particles: (1) heavy charged nuclear fragments that travel only very short distances and hence do not contribute to a lateral enlargement of a pencil beam, (2) secondary protons, and (3) neutrons.

The scattered or knocked-out, relatively high energy, protons from the second of the above categories also contribute to the tails of a pencil beam's lateral dose distribution. These protons emerge from the collision at a small but not negligible angle to the direction of the incident protons and create a halo of dose around the beam that grows in size as the depth increases. This halo, too, can be approximated by a Gaussian distribution which further contributes to the tails of the lateral dose distribution as depicted in Figure 10.14. If, while performing absolute dosimetry, this long tail to a pencil beam's dose distribution is ignored, one may underestimate the dose by many percent (Pedroni *et al.*, 2005).

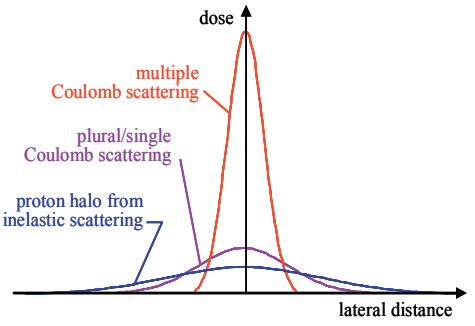


Figure 10.14. Schematic representation of the three charged particle components of the lateral profile of an initially infinitesimal pencil beam (see text).

Nuclear Interactions: neutrons

Nuclear interactions create a halo of neutrons that largely escape the patient without further interaction, but are responsible for a low dose in and outside the beam that may, for example, have consequences for secondary cancer production – especially in children or the fetuses of pregnant women (Schneider, 2002; Hall, 2006).

Pencil beam broadening due to material upstream of the patient

Broadening of a pencil beam within the patient, as discussed above, is unavoidable. Protons suffer the same type of interactions in material upstream of the patient and the amount and composition of this material is, to some extent, under one's control. Scattering in upstream material is exacerbated by any drift path (i.e., gap) between the material and the patient. A drift path allows the beam to expand after being scattered, as shown schematically in Figure 10.15. The extent of upstream scattering depends on: the method of beam formation (i.e., scanning or scattering); the method of energy control; the use or not of double scatterers (see below); the location of the aperture if any; the path in air after the last upstream material (Urie *et al*., 1986b), and so forth.

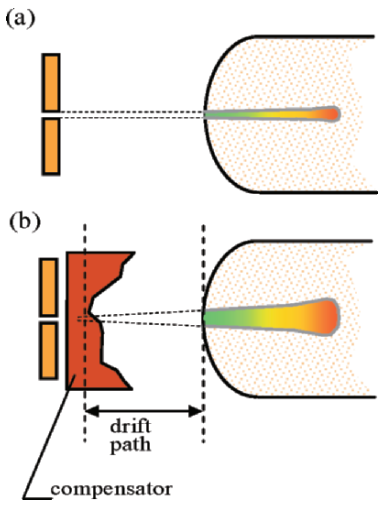


Figure 10.15. Lateral spreading of a pencil beam: (a) without, and (b) with upstream material in the beam.

One generally tries to minimize the amount and composition of any upstream material (materials of low atomic number scatter less) so as not to degrade the beam penumbra any more than necessary. In addition, the location of any material is important. As a general rule of thumb, one tries: (1) to locate the aperture close to the patient so as to reduce the size of the penumbra – but not too close so as to avoid superficial hot spots from protons scattered off the aperture edges (see below); (2) to locate material upstream of the aperture as far from the aperture as possible; and (3) to locate material downstream of the aperture as far from the aperture as possible, i.e., as close to the

patient as possible. All this represents a juggling act and a suitable compromise has to be made in practice.

Broad beams

A broad beam of protons may either be formed physically from a series of actual pencil beams, as in beam scanning as discussed below, or may be produced by scattering. Figure 10.16 shows a lateral profile of a broad beam, generated by passive scattering techniques.

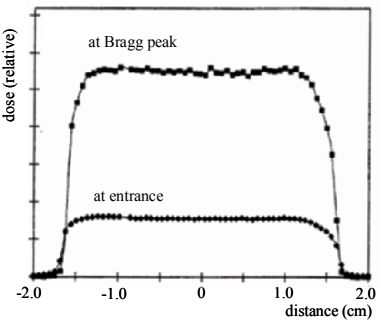


Figure 10.16. Lateral dose profile of a 160 MeV broad beam at the entrance and at the top of the Bragg peak. The beam was formed using the double scattering process. Figure courtesy of B. Gottschalk, HCL, USA.

For large proton penetrations (e.g., ≥ 20 g·cm^{-2}), the lateral penumbra near the end of range is dominated by scattering in the target material. By contrast, for small proton penetrations (e.g., ≤ 8 g·cm^{-2}), the lateral penumbra is usually dominated by blurring effects that occur upstream of the patient as a result of finite beam size, scattering in upstream material and so forth. In between, the two effects are comparable.

Typically, where multiple Coulomb scattering in the patient predominates, the least sharp penumbra is at and near the end of range and is approximately equal to a bit more than 3% of the range. Thus, a beam penetrating 15 cm could have a lateral penumbra (80%–20%) of about 5 mm. In practice, because of upstream scattering, it is more likely to be about 6 mm. This compares favorably with the penumbra of a linac-produced X-ray beam that is typically 6 to 9 mm (see Figure 4.18 in Chapter 4). At depths above about 20 cm, the proton penumbra becomes greater than that of a high energy photon beam.

ALL THINGS CONSIDERED…

Figure 10.17 sums up all the effects we have been discussing. In it are indicated a number of points throughout and outside a broad proton beam. The dominant contributions to the dose at those points are shown in the panel on the right hand side. I recommend, as an exercise, covering the panel on the right hand side of the figure, and trying to identify the principal contributions for yourself.

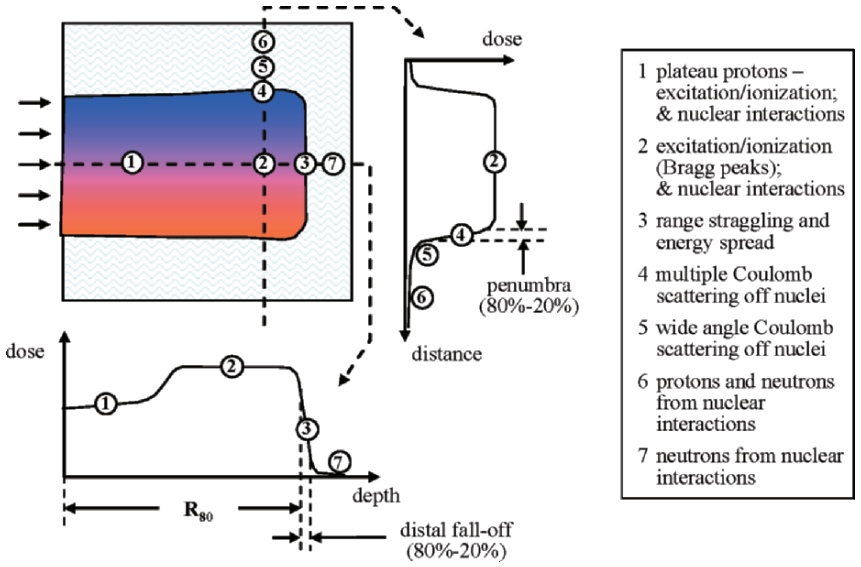


Figure 10.17. Synopsis of the principal contributions to the dose at various points within and outside a broad proton beam.

PROTON THERAPY: ACCELERATOR AND BEAM DELIVERY

Until the last decade, proton beam therapy was only available in a few physics laboratories. Only recently have purpose-built facilities been built in hospitals. A typical plan of the treatment level of such a proton medical facility is shown in Figure 10.18. For greater efficiency, several treatment rooms are usually served by one accelerator.

Particularly in the hospital setting, a significant hurdle is posed by the large size of the equipment. The size is primarily due to the high magnetic rigidity of therapeutically useful protons, which implies the need for relatively large electromagnets to transport the beam. However, the large size of proton treatment equipment does not have to imply a much greater intrinsic complexity than, say, a conventional linac. The two machines have very similar sub-systems as I'll point out in what follows. Also, in terms of operations, the two should be quite similar. Both should be "push-button" machines, not requiring dedicated operators. The proton machine is primarily different in that additional controls are needed to deal with the additional degree of freedom, namely the depth dimension. (It has to be admitted, however, that current machines do not yet achieve these goals.)

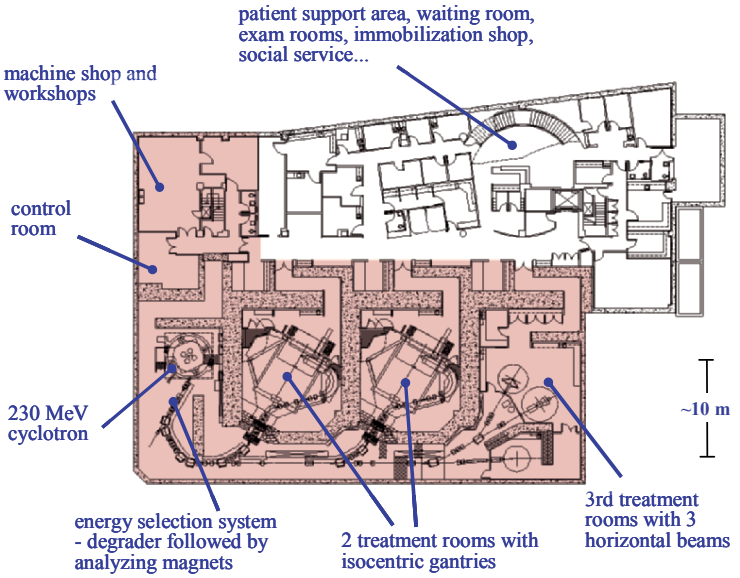


Figure 10.18. Layout of a typical proton facility. This is the treatment floor plan of the Massachusetts General Hospital's system (Boston, USA). The technical area is colored.

Accelerator

The accelerator is the "engine" of the facility but, just as in an automobile (or in a conventional linac), from the point of view of space, cost, and complexity it represents only a modest fraction (~20%) of the whole system. Relatively low beam currents, of the order of tens of nanoamperes coming into the beam spreading device, are required for radiation therapy. Cyclotrons and synchrotrons have been used, and linear accelerators have been considered. The specifications that drive the choice of accelerator are, on the one hand, the general ones of safety, reliability, and ease of operation and maintenance and, on the other hand, the requirements of the beam-spreading technique.

The required energy of the protons reaching the patient differ from patient to patient, and within the delivery of a single field. Thus, it is highly desirable to have a variable-energy proton source. A synchrotron provides energy variation very simply by extracting the protons when they have reached the desired energy. A cyclotron is

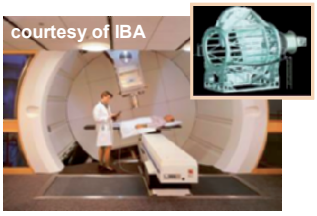
generally a fixed energy machine and, to produce protons of less energy, the energy of the protons must be adjusted downstream of the accelerator. This is done with a variable-thickness degrader, which interposes varying amounts of material in the beam, thus altering its residual penetration. The process of reducing the energy in this manner also scatters the beam (via multiple Coulomb scattering in the degrader) by a significant amount and leaves the beam with a spread of energies due to range straggling. This spreading in both angle and energy is repaired using collimators and bending magnets which act as a spectrometer, picking out a narrow band of energies with adequately small size and divergence of the resultant proton beam. The process of energy degradation is quite inefficient; in the extreme as much as 99% of the protons may be stopped in collimators and thus be "lost" from the useful beam. As a result: (a) the cyclotron must be capable of producing substantially more intense beams than a synchrotron; and (b) additional shielding is required to shield against the neutrons produced by the lost protons. Generally, these neutrons are produced far from the patient and, with appropriate shielding, do not contribute significantly to the neutron dose that he or she receives.

The time structure of the beam from the two accelerators is also different. A sector-focused cyclotron produces a virtually continuous beam, whereas the synchrotron delivers its protons in pulses, usually of a few seconds duration and with a few seconds dead time in-between the pulses. The pulse structure of synchrotrons is a complicating factor for the implementation of repainting and beam gating (see below).

An active debate goes on between proponents of the two types of accelerator. I will not jump into this debate; both can do the job.

Beam-transport system

Protons must be brought from the accelerator to the treatment delivery device. This is done using magnets to guide the beam using the same principle as allows an electric motor to work, namely the fact (discovered by Faraday in 1821) that a moving charged particle experiences a lateral force when moving through a magnetic field. Protons, since they have to be transported long distances and need magnetic lenses to keep the beam size small enough, use many magnets, linacs generally only one.



Treatment delivery device: gantry

Because proton beam therapy originated in physics laboratories, for decades only fixed horizontal beams were available. These are still used for specialized treatments such as those of ocular melanomas, but isocentrically rotating gantries have now become the beam delivery device of choice. In photon linacs, the entire system, except for the supporting power supplies and rf amplifiers but including the accelerator and beam-transport system, is packed into the gantry. With protons, the accelerator and beam transport are generally separate[3] and the gantry's function is: a) to allow the beam to be directed towards the patient from any direction in the plane of rotation; and b) to carry the beam delivery system, of which more below.

Two types of gantry have been developed: a large-throw gantry with a diameter of the order of 10 to12 meters; and a compact gantry with a diameter in the range of 4-7 meters, such as the prototype realized at the Paul Scherrer Institute in Switzerland and now in clinical use there. Both consist of a large mechanical structure which is supported on rollers or roller bearings and can rotate by somewhat more than 360°. The structure supports a series of magnets – really just an extension of the beam-transport system – and the beam delivery system, which is discussed below.

Additional flexibility in beam direction and positioning capability is provided by a treatment couch that has, in proton therapy, evolved to have all six degrees of freedom of movement; in addition to the linac couch's three directions of translation and rotation in the horizontal plane about isocenter, pitch and roll motions are provided. These additional degrees of freedom allow for easy correction of the patient's orientation without requiring large motions of the gantry and couch.

Tumors frequently abut, or are very close to, critical normal tissues. This leads to a very tight requirement on overall beam-pointing accuracy of 1 mm or better – which requires that the gantry and patient positioner both be extremely reproducible and that the patient

[3] Although, at the time of writing, gantry-mounted accelerators are being considered for "single room" proton therapy.

position be very tightly controlled. As a consequence, major technical challenges are: to provide sub-millimeter mechanical precision on a moving system with a weight of the order of a hundred tons or more; and to guarantee beam shape invariance and positional stability within a few tenths of a millimeter during rotation of the gantry.

BEAM DELIVERY SYSTEM: SCATTERED BEAMS

Target volumes typically range in size from a few milliliters to several liters. As a consequence, the pencil beam emanating from the accelerator and transported through the gantry, with its small diameter and small extension of the Bragg Peak in depth, generally needs to be spread out, both laterally and in depth. These goals are accomplished in what is often referred to as the *nozzle* which is placed at the end of the beam transport element of the gantry and constitutes the *beam delivery system*. There are two main approaches for shaping the beam laterally, scattering and scanning (see below).

The historical approach, which is still in wide use today – in all but one proton medical facility at the time of writing – spreads the beam out laterally by a passive[4] scattering technique, while spreading in depth is accomplished by a "range modulator." Figure 10.19 shows the basic elements.

Lateral enlargement of the beam

The proton pencil beam is spread out laterally by interposing scattering material so as to produce a broad beam with a homogeneous flux of particles throughout the solid angle covering the tumor. This lateral dispersion can be done most straightforwardly with a single piece of scattering material (usually chosen to have high atomic number so as to minimize the energy loss for a given amount of scattering), but the efficiency of this process is low since, due to the Gaussian shape of the scattered beam, no more than about 10% of

[4] Because the lateral spreading of the beam is performed by a static piece of material, systems that use scattered beams are often referred to as providing passive beam delivery. Since the spreading in depth of such beams is usually performed by a rotating range modulator, there is usually a dynamic element to the systems and the use of the term "passive beam delivery" is not strictly correct.

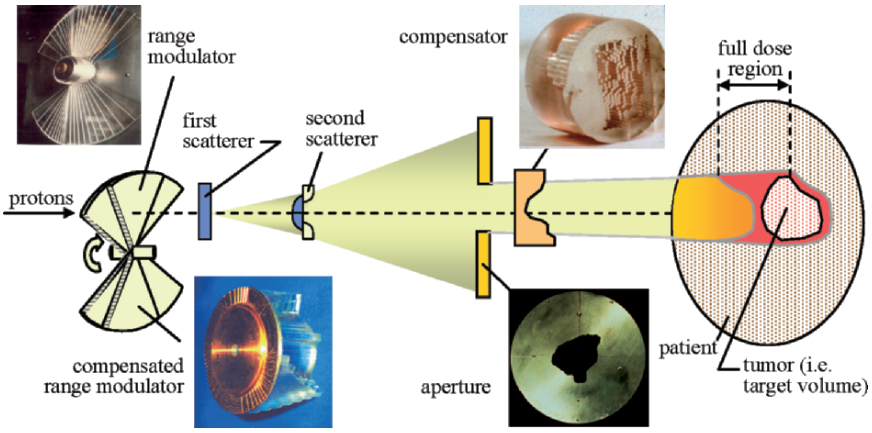


Figure 10.19. Schematic diagram, not to scale, of a passive scattering nozzle (see text). *N.B.* monitoring devices are omitted in this diagram.

the protons lie in the near-flat region at the center of the scattered beam, and are therefore useful.

Double-scattering systems are now in widespread use. These can be of several designs. In perhaps the most sophisticated approach, the first scatterer is a simple piece of material of uniform thickness, usually of high atomic number (since this minimizes energy loss for a given degree of scattering), and the second scatterer, located some distance downstream from the first, is shaped so that it preferentially scatters the center of the beam more than the outside. This system transmits a substantial fraction of the beam (up to about 45%) which is uniform enough to use for treatments (Gottschalk, 2004). In an additional refinement, the range modulator (see below) and first scatterer can be combined, using both high and low atomic number materials (see inset in the lower left of Figure 10.19), in order to achieve a constant level of energy loss throughout the useful beam (Gottschalk, 2004).

Double scattering systems have a couple of disadvantages that must be overcome: (1) because they spread out the beam in two separated scatterers, they produce a beam with a much larger effective source size than is produced by a single scattering system and consequently have a larger penumbra; and (2) to obtain a flat dose distribution, the beam must be very well centered on the contoured second scatterer.

Both scattering methods have their place in proton beam therapy. In particular, where not-too-large fields with excellent penumbral quality are needed, a singly-scattered beam is to be preferred. For the largest fields, where efficient use of protons and reduction of secondary dose from absorbed protons are desirable, doubly-scattered beams are usually used.

Tailoring the beam in depth: the range modulator

A *range modulator* is a rapidly rotating device that interposes a sequence of different thicknesses of material into the beam for varying durations, thus delivering a sequence of Bragg peaks of incrementally different ranges and weights. The required characteristics of a range modulator depend on the size and depth of the tumor; for a given patient and beam one needs to select from a library of prefabricated range modulators. In most recent facilities, a number of these are mounted on a motorized carousel, which is included in the nozzle and allows for automated insertion of the desired range modulator into the beam.

Ridge filters have been used instead of the type of range modulator just described. These are absorbers with multiple "ridges" shaped so as to transmit just the right spectrum of proton energies so as to achieve the desired depth–dose distribution.

Passive scattering has the limitation that the extent in depth of the high dose region of the SOBP is inherently uniform everywhere within the field. Thus, the extent in depth of the SOBP is set by the maximum extent in depth of the target volume, causing unnecessarily high dose upstream of any thinner portions of the target volume as illustrated in Figure 10.19.

Tailoring of the depth of penetration: compensators

One wishes the distal edge of the proton beam, with some safety margin added, to coincide precisely with the posterior surface of the target volume. This conformity is achieved through the use of patient-specific range *compensators* – devices that are thin where a large beam penetration is desired, and thick where little beam penetration is desired. Compensators are generally made of low atomic number material such as plastic to minimize the amount of scattering that they cause. Compensators need to be tailored to the individual patient and designed to accommodate a number of uncertainties. The discussion of how this may be done is deferred to Chapter 11.

Achieving a sharp penumbra: apertures

A sharp boundary to the dose distribution in the lateral direction is generally desirable in order to spare, to the extent possible, normal tissues located lateral to the beam and outside the target volume. A sharp penumbra is achieved by interposing individually shaped patient-specific apertures – just as for photon therapy (see Chapter 4). However, there are a couple of differences in the case of protons.

The first difference that scattered proton beams have, relative to photons, is that, when double-scattering techniques are used to spread the beam out, one is left with a very large effective source size, of the order of centimeters in diameter, as opposed to a photon linac or a single-scattered proton beam where the source size is of the order of millimeters. One must therefore move the aperture as close to the patient as possible in order to minimize the penumbra (as illustrated in Figure 4.16 of Chapter 4).

The second difference is that apertures are prone to produce a small contamination of low energy protons coming from their edges, which can result in somewhat higher dose being delivered to mainly fairly superficial tissues. As shown in Figure 10.20, the contaminants come from a very small peripheral strip of the order of a millimeter or so wide, within which protons may be scattered out of the aperture material and towards the patient – of course, with reduced energy due to their having traversed some of the aperture material.

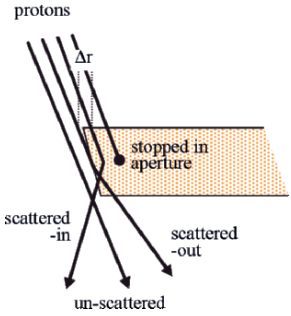


Figure 10.20. Illustration of the source of low-energy protons produced at one edge of an aperture (see text).

The ratio of the net flux of edge-scattered protons relative to the flux of protons passing through the aperture is small – of the order of $2\pi r\Delta r/\pi r^2$ where Δr is the width of the strip, and r is the radius of the aperture. This computes to about 5% for an 8 cm diameter field, and linearly less for larger fields. Although this is a small number, the scattered low energy protons can lead to undesirable hot spots in the shadow of the aperture edge. To diminish their effect: (1) the aperture edge should be tapered to within about 1° of the beam's divergent edge in order to present as small a strip of material as possible to the incident protons; (2) it is useful to locate the

compensator, when there is one, downstream of the aperture so that it can absorb at least some of the low energy protons; and (3) one likes to allow a drift path between the aperture and the patient so that the low energy contaminants can spread out, producing a lower dose, but over a greater area. A compromise has to be arrived at in this last approach as it is in direct opposition to the desire to have as small a drift path as possible to avoid degrading the penumbra.

Proton apertures are usually made of a material such as brass, which allows them to be relatively compact, while producing a lower neutron background than an even denser material such as, say, lead or tungsten.

Multi-leaf collimators, with their potential to be remotely controlled, are potentially useful types of aperture, just as for photons (see Chapter 4). However, their generally bulky size is in conflict with the desire to keep the aperture as close to the patient as possible so as to have as sharp a penumbra as possible, and they have not found widespread use in proton beam therapy as of the time of writing.

BEAM DELIVERY SYSTEM: SCANNED BEAMS

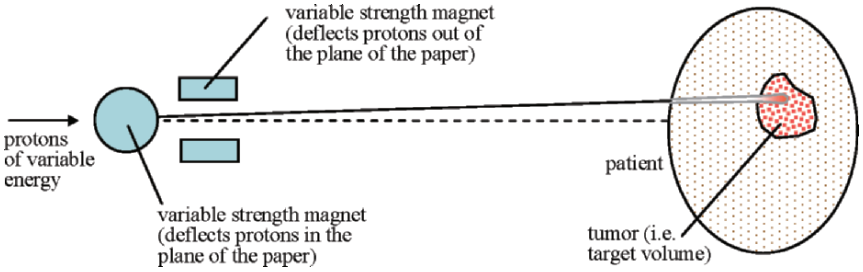


Figure 10.21. The basic elements of a scanning nozzle (see text). *N.B.* monitoring devices are omitted in this diagram.

Figure 10.21 shows the basic elements of a scanning nozzle in which the dose can be spread laterally, using a pair of magnets whose excitations can be rapidly varied, and can be spread in depth by making energy changes upstream. Scanning sweeps a finite pencil beam through the target volume in a predetermined pattern, laying down, as it were, Bragg peaks (or "spots") wherever they are needed, each spot being delivered with whatever intensity is desired. Scanning can be accomplished by either magnetic or mechanical means, or by a combination of the two.

The scan program may either be implemented by a sequence of static pencil beams in which, after one pencil beam has delivered its dose, the delivery of protons is interrupted, the pencil beam is moved to its next position in the sequence, and the delivery of protons is resumed. This approach is referred to as *spot scanning*. It much resembles the "step-and-shoot" approach used in photon IMRT and has the advantage that the delivery of each "spot" is quite similar from the controls point of view to the delivery of a broad beam. Alternatively, pencil beams of a given energy may be swept through a pre-determined pattern (e.g., a raster scan as used in video monitors) and the intensity of proton delivery varied as needed during the scan.

Scanned beam delivery has several advantages.

1. It can "paint" any physically possible dose distribution.
2. It uses protons very efficiently as compared to passive scattering techniques in which more than 50% of protons have to be "thrown away."
3. It generally requires no patient-specific hardware[5] – as a consequence of which a treatment fraction consisting of several differently directed beams can be delivered quickly, without the need for the therapist to enter and leave the room between beams to change out the aperture and compensator.
4. The neutron background is substantially reduced as a result of points (2) and (3).
5. And, most important of all, scanned beam delivery allows the implementation of IMRT with protons – termed *intensity-modulated proton therapy* (IMPT).

Scanned beam delivery also has some disadvantages. Foremost among them are the following.

1. The need for heightened safety measures due to the dire consequences of instrumental or control system failure, which could result in a high intensity pencil beam lingering on the patient, rather than moving on to the next position in the scanning program.
2. The need to overcome interplay effects induced by organ motion.

These matters are further discussed below.

---

[5] Although, in some situations, it would be desirable to improve the penumbra through the use of field-edge trimmers.

The versatility of scanned beam delivery, its ability to deliver IMPT, and the fact that it makes it possible to treat a patient without patient-specific hardware, will, I feel sure, result in its largely replacing scattered beam delivery within a few years.

The only clinically active scanning system for protons as of the time of writing is the compact spot-scanning gantry at the Paul Scherrer Institute in Switzerland. There, lateral beam spreading is performed longitudinally via magnetic deflection of the beam and transversely by moving the patient table. The modulation of beam range is realized by dynamically changing, typically in about 50 ms, the amount of material in front of the patient. A scanning system for heavy ions has been developed at GSI in Germany. It features magnetic scanning in both lateral directions, and dynamic variation of the extracted beam energy from their synchrotron.

Figure 10.22 shows a set of pencil beam dose distributions, such as would be used for beam scanning, covering a range of energies. The desirable pencil beam size for treating other than superficial tumors is about 5 to 8 mm full-width at half-maximum. Smaller beams are difficult to obtain due to multiple scattering in the equipment and in the patient. In practice one typically needs to deliver from about 1,000 to 30,000 discrete Bragg peaks (or, "spots") to build up a broad beam, depending on the volume of the target. Since at any given time the full beam intensity is being delivered to a rather small volume, safety is a major concern. Redundant measuring systems, two independent computers, and fast beam-abort systems in case of malfunction of a component, are all required. While scanning used to be considered much more complex and, therefore, more liable to error than passive scattering, a comparison of Figures 10.19 and 10.21 somewhat softens this judgment (although, it has to be admitted that the omitted monitoring devices are quite a bit more complex in the case of beam scanning).

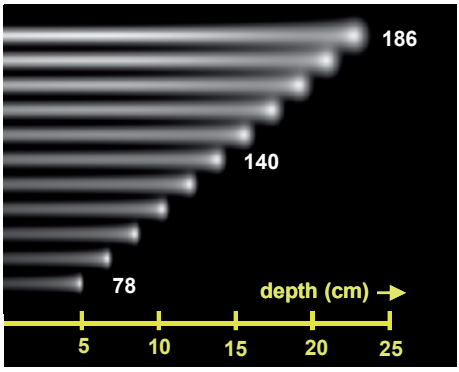


Figure 10.22. Calculated (3mm sigma) pencil beams of protons, ranging from 78 to 186 MeV. Figure courtesy of E. Pedroni, PSI, CH.

Interplay effects due to organ motion

The major problem with scanning is its sensitivity to organ and tumor motion during the time a beam is being delivered, primarily due to respiration. Organ motion can markedly affect the dose distribution because of what are termed *interplay effects* (Bortfeld *et al*., 2002). This term describes the possible interplay between the motion of the scanned pencil beam, and the motion of, say, a cell (Goitein, 2005) within the target volume. A given cell can either move so that it is outside a pencil beam when it should be within it, resulting in a lower than desired dose, or can linger within a pencil beam as it moves, resulting in a higher than desired dose. Figure 10.23 explains graphically, although in an exaggerated manner, how this variation in dose, which I term *dose mottle*, comes about. On the positive side, the average dose within the target volume is essentially unchanged as one is delivering the same total energy no matter what movements take place.[6]

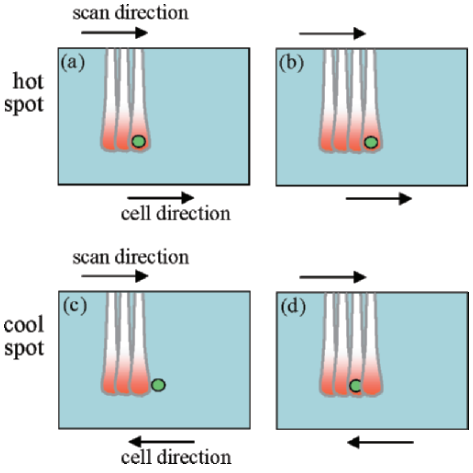


Figure 10.23. Schematic illustration of how dose mottle arises. (a) and (b): cell irradiated by the third pencil beam moves right, into the way of the fourth pencil beam, and so receives near-double dose. (c) and (d): cell initially outside the third pencil beam moves left, avoiding being irradiated by fourth pencil beam, and so receives near-zero dose.

There are two ways to deal with interplay effects, and both are needed in situations where the amplitude of motion is more than a couple of millimeters or so. The first approach is to gate

[6] One can potentially have a problem any time the period of the beam application is comparable to the period of organ motion. This is why, with scattered beams, the range modulator is generally designed to rotate rapidly (hundreds of Hz); its period is then much shorter than any significant organ motion and, hence, interplay effects are averted. In scanned beam delivery, because one scans in three dimensions, at least one motion is pretty much guaranteed to have a period comparable to that of the respiratory cycle, for example, and thus give rise to interplay effects.

the beam according to the patient's respiratory cycle – an approach which has already been described in Chapter 7.

The second way to deal with interplay effects is to "paint" the dose distribution not once, but many times, with a correspondingly lower dose per painting. Provided repainting is done at times comparable to or longer than the period of respiratory motion (~5 s), the dose mottle will tend statistically to average out to an acceptable level. Typically, some 10 repaintings are desirable. In principle, the higher-weighted spots need repainting more often than the lower-weighted spots. However, this does not mean that, in practice, only the higher-energy layers need to be repainted. Due to curvature of the target volume, a number of different energy layers can include high-weighted spots. If one wishes to keep the average irradiation time per beam down to about one minute or so, one must be able to execute one painting in approximately one tenth of that time, i.e., in approximately 6 s. Such a short time can be technically difficult to achieve, but it is within the scope of current technology.

Beam wobbling

Beam wobbling is a variant of beam scanning in which the scanned pencil beam is considerably larger than that used in normal scanning while still being smaller than the size of the full field to be treated. The downside of beam wobbling is that the pencil beam size is too large to give a sharp penumbra, or to compensate for inhomogeneities, or to deliver IMPT. One has to use apertures and compensators in wobbled beams, thereby losing many of the advantages of scanned beams.

Why then would one want to do beam-wobbling? There are a few reasons. Perhaps most importantly, motion artifacts are quite a bit less in wobbling than in scanning; dose mottle is reduced by roughly the ratio of the pencil beam widths used in the two techniques, which can be at least a factor of five. Thus, in situations in which organ motion is so substantial that it is judged that gating plus repainting would still not reduce dose mottle sufficiently, beam wobbling may be the answer – augmented by respiratory gating and repainting when possible. Then, wobbling can easily produce large fields (which may be hard to deliver using scattering approaches) and can be quite a bit more efficient than scattering, thus allowing large uniform fields to be delivered with fewer protons. And, finally, fewer protons may be lost on the collimators, thus reducing secondary neutron radiation.

Current status of beam scanning

Proton beam scanning has, at the time of writing, only been in substantial clinical use in one center (PSI, Switzerland) – and, for heavy ions, in only one center (GSI, Germany). In contrast, tens of thousands of patients have been treated using scattering, with generally excellent results. Thus, beam spreading by scanning is in its infancy. However, the ability of scanning to deliver intensity-modulated proton therapy, together with other potential advantages, means that beam scanning is destined to be used widely in the future – although it may not entirely replace scattering.

BEAM CONTROL

Monitoring and dosimetry

I have completely glossed over the not-inconsiderable problem of monitoring the beam to assure the desired dose distribution is delivered extremely reliably and safely. Figures 10.19 and 10.21, for example, leave out all beam monitoring devices. Suffice it to say that the beam line, the gantry, and the nozzle all need to be fully instrumented to assure the correct intensity, position and angle of the beam at any location. Just as for conventional photon linacs, the dose delivered to the patient is of highest concern. In general, redundant monitors and real time beam checking are required. The beam monitors add not inconsiderably to the scattering of the proton beam and, hence, are a potentially limiting factor in achieving a small beam spot.

Control and safety systems

It goes without saying that, just as for conventional photon linacs, comprehensive, well-engineered, safety and control systems are required. I only want to mention that the development of these systems is an enormous task, which is very often underestimated by constructor and client alike.

DOSIMETRY

I mention only briefly the matter of performing dosimetry for proton beams; (ICRU72, 2007) should be consulted for a full treatment of this subject. The recommended protocol for absolute beam dosimetry at a proton facility is presented in IAEA (2000).

Absolute dosimetry

There are three main classes of absolute dosimeter, namely:

Calorimeters These devices measure the amount of heat generated per unit mass of absorbing material – which is usually either graphite or pure water. This is very close to a direct measurement of dose; the main correction is for the so-called "heat defect" which estimates the small fraction of the energy which goes, not into heat, but into induction of chemical changes in the absorbing medium.

Faraday cups The Faraday cup measures the total charge deposited in a block of conducting material when protons stop entirely within the material. Since the proton charge is very accurately known, one can directly ascertain the number of protons that stopped in the material. Armed with this information, and knowledge of the stopping power of the protons coming in to the Faraday cup, one can immediately compute the dose that would be delivered to material placed just in front of the device. The main correction is for the change in collected charge due to charged particles such as electrons which escape the stopping material by being scattered backwards and, to a lesser extent, for the change in charge due to absorbed secondary particles emanating from material in front of the Faraday cup.

Ionization chambers Ionization chambers consist of a pair of electrodes between which is sandwiched a known amount of gas such as air. The gas is ionized by the radiation traversing it, and an electric potential applied across the electrodes separates the positive ions and the electrons which drift towards opposite electrodes where they are collected and the total charge measured. Ionization chambers may intercept the whole beam, or may be used to measure the dose in a small volume. In the former case, they use large diameter parallel plates which integrate the dose across the full beam. For measurements of small volumes, ionization chambers can be made quite small, containing perhaps a fraction of a milliliter of gas. They are often cylindrical in shape with a wire forming the central electrode, They may also feature a parallel plate geometry which is useful for measurements of dose close to the entrance surface. Converting the current coming from an ionization chamber into dose requires knowledge of the mass of gas contained in the cavity, and the so-called "w value" which is the amount of energy required to ionize an atom of the gas.

Historically, there has been some question about the reliability of the Faraday cup (Verhey *et al*., 1979) and current practice is to use an ionization chamber, calibrated in a ^{60}Co photon beam and using a w-value which is determined experimentally, most usually by inter-comparison with a calorimeter. There is now international agreement (ICRU78, 2007) to use the protocol described in IAEA (2000) to convert ionization chamber measurements to proton dose.

Relative dosimetry

Relative dosimetry is used for a number of different purposes.

Beam line monitoring

A number of monitors are required between the accelerator and the beam delivery system, primarily to measure the beam intensity and position along the beam line. Typically, large-area parallel-plate ionization monitors, are used. For beam position measurements, the collecting electrode is usually divided into several sections (e.g., four quadrants), the output of each of which is separately measured. By looking at the ratio of outputs from opposing segments or combinations of segments, once can deduce how well the beam is centered on the monitor.

Machine output

While measurement of the absolute dose is essential for radiation therapy, the monitor which controls the delivery of the proton beam is usually a relative monitor – namely, a parallel-plate ionization chamber covering the entire beam. This device is calibrated against an absolute dosimeter, placed within a block of near tissue-equivalent material and irradiated under standard conditions, on a regular basis.

Dose distribution measurements

Routinely, the dose distribution throughout a delivered beam has to be measured. This is usually done using a relative dosimeter and normalizing the distribution to a calibration point where the dose is ascertained using an absolute dosimeter.

Dose distributions may be measured using the small ionization chambers just described, diode detectors, a scintillation screen viewed by a CCD camera, or film. Diodes have an LET dependence which makes them more suitable for scanning lateral dose distributions than dose distributions in depth where they tend to disagree with the preferred ionization chamber measurements by perhaps 10% at the

top of the Bragg peak. In particular, the very small effective collecting volume of diodes make them particularly suitable for measuring beam penumbra. Film, if used alone, measures fluence rather than dose. It must be placed in close optical contact with a scintillation screen for it to register dose.

IN CONCLUSION

In this chapter, I have tried to emphasize the clinically relevant physics of protons and how they are delivered to the patient. An understanding of the underlying mechanisms and techniques makes the use of any tool much more secure, and offers opportunities to move beyond accepted wisdom. In the next chapter, I discuss the application of proton beams tailored to the individual patient in the clinical setting.

11. PROTON THERAPY IN THE PATIENT

The previous chapter dealt with aspects of proton beam therapy which are more or less independent of the patient. That is, it discussed matters that do not depend on the patient's geometry nor on the treatment planning aims. This chapter addresses the more clinical side of the problem; how one tailors proton beams to the patient, and how one plans a treatment using protons.

[*]At the time of writing, substantial changes are taking place in proton beam therapy. For one thing, proton beam therapy has moved from the physics research laboratory, where it was sequestered for several decades, into the clinic and there has been a rapid growth in the number of proton medical facilities worldwide. On the technical side, practitioners are beginning to emphasize pencil beam scanning more, and scattered beam delivery less. The reasons for this have already been alluded to in Chapter 10. For uniform-beam planning, scanned beams: (1) offer better sparing of tissues upstream of the target volume because the depth of modulation can be varied throughout the field on the basis of the extent in depth of the target volume; (2) can in most cases be delivered without patient-specific hardware, thus allowing treatments to be given more quickly, and to be adapted to changing circumstances more easily; (3) can create non-uniform beams and so be able to deliver intensity-modulated proton therapy (IMPT); and (4) produce less neutron background radiation than scattered beams. The downside is that one has additional issues to deal with, related to patient and organ motion, as discussed in Chapter 10. While technically challenging, motion does not pose a fundamental limitation in all but a few instances – for which wobbled beam delivery, described in Chapter 10, can be used.

I propose to begin with a discussion of a number of topics which apply equally to all forms of beam delivery, although the technical implementations may differ, depending on whether the beams are delivered with scanned pencil beams or scattered or wobbled beams. The first topic is the influence of inhomogeneities on a beam's dose distribution.

INHOMOGENEITIES

A patient's tissues are highly inhomogeneous both in chemical composition and density. Such inhomogeneities affect the dose distribution of proton beams which must therefore be designed to take the patient's anatomy into account.

---

[*] Some of the material in this chapter is adapted, with permission, from the article "Treating Cancer with Protons" which appeared in the September 2002 issue of Physics Today (pp. 45-50) by Goitein M., Lomax A.J. and Pedroni E.S. A good source of information concerning proton beam therapy can be found in ICRU78 (2007), from which portions of this chapter have been taken with the permission of the Oxford University Press.

Dose perturbations due to simple inhomogeneities

Inhomogeneities affect the dose distribution of a proton beam in two principal ways: (1) they affect the penetration of the protons distal to an inhomogeneity; and (2) they cause dose perturbations due to differences in scattering between adjacent regions differing in density and/or composition. The influence of inhomogeneity is discussed in the context of four scenarios, the first three of which are illustrated in Figure 11.1:

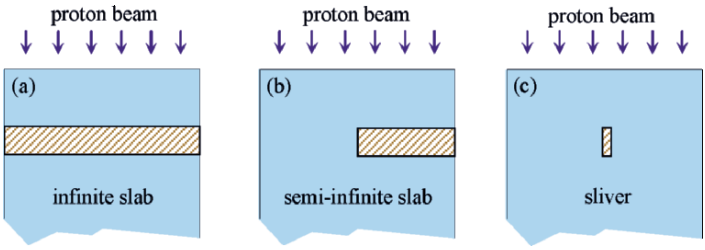


Figure 11.1. Three prototypical cases of simple inhomogeneities: (a) infinite slab intercepting a proton beam; (b) a semi-infinite slab; (c) a sliver.

Uniform infinite slab intercepting all of a proton beam

Photon and proton beams are very differently affected when they traverse a uniform slab of material whose density and/or composition differs from that of the surrounding medium. A photon beam's intensity, and hence dose, is reduced (for a higher density inhomogeneity) by typically a few percent – the amount of the reduction depending on the slab's thickness and composition. Under the same circumstances, a proton beam's intensity just after the inhomogeneity is virtually unchanged, but its penetration (i.e., range) beyond the slab is strongly affected – the amount of change depending on the slab's thickness and composition. The difference is schematically illustrated in Figure 11.2a.[1] Illustrating this point graphically, Figure 11.2b shows the results of an experiment in which a radiograph was taken behind a lamb chop immersed in a water tank

[1] I always remember the reaction of a colleague to whom I showed this picture. I had expected him to express great concern that an inhomogeneity could cause protons to underdose a distal part of a tumor. "Well," he remarked, "that's great. Protons don't lose intensity behind an inhomogeneity."

and exposed to either diagnostic energy X-rays (top image) or protons (lower image). The proton beam energy was selected so that the film would be situated in the falling distal edge of the spread-out Bragg peak, where dose falls rapidly as the thickness of upstream material increases. This stratagem accentuates range differences. One sees the much higher contrast of the proton image as compared to the X-ray image. This is due to the ranging out of protons in the shadow of high-density bones and to the greater penetration of protons behind the lower density fatty regions within the specimen.

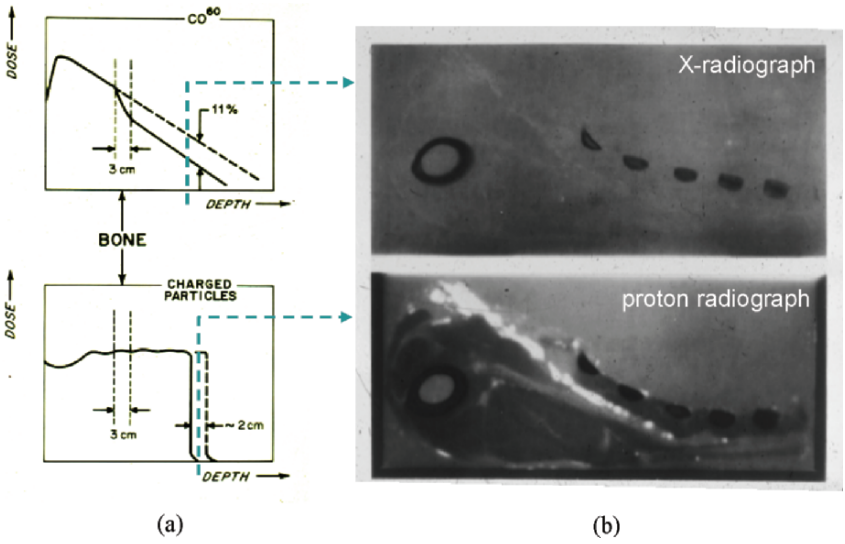


Figure 11.2. Influence of inhomogeneities on the depth dose distributions of photon (*top*) and proton (*bottom*) beams: (a) depth dose curves of a slab inhomogeneity; and (b) radiographs of a lamb chop in a parallel-sided water tank (courtesy of A.M. Koehler, HCL, USA).

If ever there were an example of the saying that a picture is worth a thousand words, this is it. The images of Figure 11.2 have lodged in my mind throughout the years, forming a subliminal reminder of the importance of inhomogeneities in charged particle therapy.

Semi-infinite slab intersecting part of a proton beam

What happens when a slab of material of a density different from that of the surrounding medium is interposed into only part of the beam cross-section? Away from the interface between the two media, the beam penetration is altered in the shadow of the inhomogeneity just as for the case of a fully intersecting inhomogeneity, and is

unchanged in the region not shadowed by the inhomogeneity. However, in the shadow of the interface region, the dose is additionally perturbed due to the difference in the strengths of multiple scattering in the two adjacent materials. Namely, there is a dose enhancement (hot spot) on the low density side, and a dose reduction (cold spot) on the high density side (Goitein, 1978; Goitein *et al*., 1978). The geometry is illustrated in the left side of Figure 11.3 for the extreme case of a parallel beam of protons impinging on an air/plastic interface situated in air.

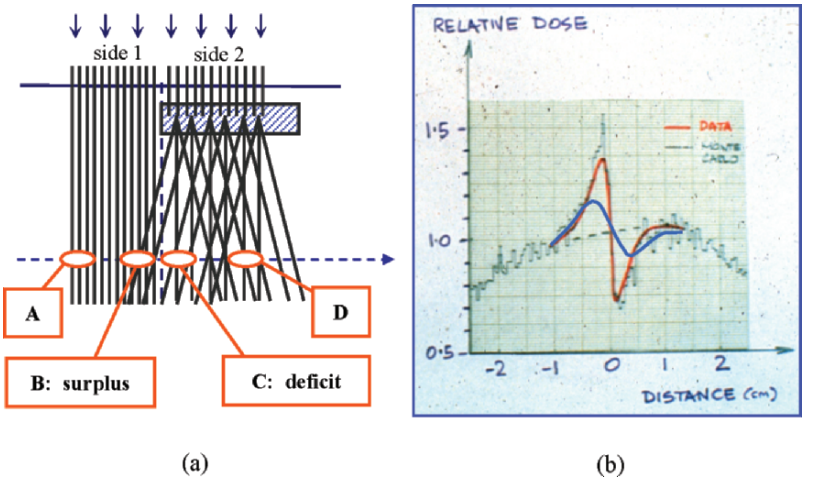


Figure 11.3: Influence of semi-infinite slab intersecting a broad proton beam, traveling in air: (a) schematic drawing (see text); and (b) measured (*red*) and Monte Carlo simulation (*black histogram*) of the dose in air 25 cm below a 2.5 cm thick half-slab of plastic. The *blue curve* shows measured data when a 1.25 cm thick infinite slab of plastic is placed just above the semi-infinite slab.

Protons on side 1 miss the inhomogeneity and travel on, unperturbed. On the other hand, protons on side 2 will be scattered by the inhomogeneity. Protons reaching the region marked A in Figure 11.3a come entirely from side 1 and will deliver a dose equal to that which would be delivered if the slab were not present. The same fluence of protons as reaches A will reach regions such as that marked D which are well away from the shadow of the edge of the inhomogeneity since, although protons have been scattered by the inhomogeneity, the net flux of particles does not change.

However, near the shadow of the edge of the inhomogeneity things are different. A region such as B will be traversed both by protons on

side 1, and by some protons which came from side 2, but were scattered into side 1 by the inhomogeneity. The flux of protons will be increased by these additional protons, and so the dose at B will be increased over the dose at, say, A. In region C, the opposite occurs. Protons from side 1, since they are not scattered, do not reach C and cannot contribute to its dose. Some protons initially from side 2 will be scattered out of side 2 into side 1 by the inhomogeneity, so that the flux of protons at C will be diminished. One says that "scattering-in" has not compensated for "scattering out." As a consequence of these effects, for an initially parallel beam of protons, the dose perturbation is as high as ±50%. Figure 11.3b presents some data illustrating this effect[2]. The dose perturbation was less than the theoretical limit of ±50% due to the fact that the proton beam was not perfectly parallel.

The perturbation is substantially modified if the beam has significant angular confusion (i.e., the protons have a finite distribution of directions at points within the beam) – such as would be induced by overlying material. For example, when an additional layer of tissue of only one-half the thickness of the tissue in a tissue: air interface is interposed above the interface, the dose perturbation is reduced to approximately ±12% as indicated by the blue curve in Figure 11.3b. If one side of the interface is not air, but rather the interface is between two materials of different scattering powers, then the dose perturbation is much reduced; in the case of a bone/tissue interface, the perturbation is reduced from ±50% to approximately ±9% (Goitein 1978; Goitein *et al*., 1978).

"Sliver" of material traversed by a proton beam

Figure 11.1c shows schematically the case of a "sliver," by which is meant an inhomogeneity which is thin and through which the beam passes parallel, or nearly so, to its long axis. The sliver is assumed to be thick in the direction normal to the paper. The prototypical example of a sliver is a thin section of bone, embedded in tissue. That this can affect a proton beam is clear from the proton radiograph in Figure 11.2b in which quite fine bone detail can be observed.

If the width of a sliver were to be large with respect to distances over which protons scatter, then it is obvious that beneath the body of the sliver the beam would be affected just as if one were dealing with an

[2] The graph was photographed directly from my notebook at a time when my photography was a less than an exact art.

infinite slab, but in the shadow of the edges of the sliver there would be a dose perturbation typical of that seen in the shadow of the edge of a semi-infinite slab. The question is, what happens when the sliver is thin? How do these perturbations add up? Does the sliver pull-back the penetration of the protons in its geometric shadow, or does scattering outside the shadow of the sliver fill in the dose behind the sliver? Figure 11.4 shows the of a Monte Carlo calculation[3] for this situation, for

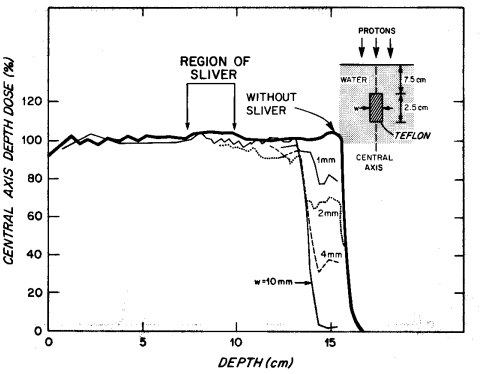


Figure 11.4. Monte Carlo calculation of the dose along the central axis of the beam through a sliver of high density material (Teflon) embedded in water for various thicknesses of the sliver. Reproduced with permission from Goitein and Sisterson (1978).

[3] A Monte Carlo calculation, in the context of a physics problem (e.g. the calculation of the dose distribution of a proton beam), is one in which a series of "histories" (e.g., of a proton traversing matter) are simulated. In each history, various physical processes are experienced (e.g., the proton interactions listed in Chapter 10) and are simulated in a computer program. Some quantity or quantities of interest (e.g., the dose) can then be estimated from the cumulative contribution of the histories (e.g., the sum of the doses deposited in a volume element from each incident proton's history). The technique gets its somewhat risqué name from the fact that the starting values for the histories, and the physical effects that are simulated, are all picked at random (by the throw of the computer-equivalent of a Monte Carlo croupier's dice) from the theoretically known distribution of possibilities. Even if the physics of the interactions is perfectly well-known, there will be statistical uncertainties in the results. The more histories that are followed, the smaller those uncertainties will be (reducing approximately as the square root of the number of histories). As a result, a very large number of histories is needed; ten million histories would not be at all unusual in calculating the dose distribution of a proton beam to ± 2% accuracy (SD), for example. This makes Monte Carlo calculations quite slow. However, they are intrinsically quite accurate – as accurate as the knowledge of the physical processes allows.

various sliver thicknesses. It is noteworthy that even a 1 mm thick sliver results in a non-trivial *circa* 20% reduction of the central axis dose in the region of the proton beam's end of range.

One needs to be able to compensate for, or at least take into account, inhomogeneities, and the results shown in Figure 11.4 are particularly alarming because they suggest the high level of spatial resolution needed. One millimeter is near the limit of resolution of CT scanners, and this means that inhomogeneities which can affect the dose distribution may escape detection or, at the least, may be inadequately measured.

Dose perturbations due to complex inhomogeneities

So far, I have focused on inhomogeneities of simple shape since they exhibit the behaviors of protons in a pure form. In practice, of course, the patient usually presents a complex pattern of inhomogeneities; this is perhaps most extreme in the region of the base of skull where protons may be directed along extended bone surfaces, or through complex bone/tissue/air structures such as the petrous ridge or paranasal sinuses. In consequence, a complex combination of range penetration perturbations and scattering-induced dose non-uniformities takes place. The results of such complex situations are very hard to calculate analytically, although the preceding discussions of inhomogeneities gives some insight into the extent of the possible dose perturbations. Monte Carlo calculations are presently the only way to get a reasonably reliable estimate in the case of highly complex geometries. To be accurate, though, a very fine calculational grid must be used.

Figure 11.5 shows the degradation of the terminal region of two different proton beams – a monoenergetic beam delivering a single Bragg peak (upper panel), and a spread-out Bragg peak (lower panel) – which was passed through a water-filled human skull. The depth dose was measured in a water tank placed downstream of, and close to, the skull. Measurements were made at three points, identified in the left panel of Figure 11.5: (A) downstream of a relatively homogeneous region of the skull; (B) downstream of a fairly inhomogeneous region of the skull; and (C) downstream of a highly inhomogeneous region of the base of skull.

These data demonstrate that the distal portion of the dose distribution can be very substantially affected by complex inhomogeneities. *The distal fall-off of both the Bragg peak and the SOBP is not simply*

shifted in range by a complex inhomogeneity. Rather, its slope can be substantially less steep and less regular. There could be both a significant underdose in a tumor, and a significant overdose in a distal-lying normal tissue if this degradation was not appreciated, or was ignored. The degradation of the falling edge of the Bragg peak is as much as ± 2 cm at point C of Figure 11.5. In a companion experiment, looking at the Bragg peak degradation of carbon ions passing through the abdomen, the degradation was even somewhat greater – which was attributed to the effects of organ motion during the long exposure needed to take the data.

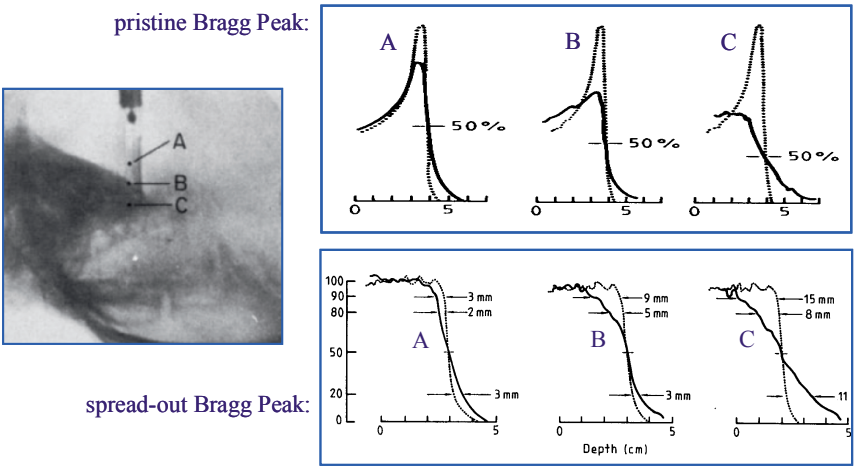


Figure 11.5. Degradation of a pristine Bragg peak (*right panel – top*) and a spread-out Bragg peak (*right panel – bottom*) passing through a water-filled human skull along paths behind three regions, A, B and C, identified in the radiograph shown in the left panel (see text). The unperturbed dose (i.e., when the skull is replaced by a water tank) is shown as a dotted line in all panels. Reproduced with permission from Urie *et al*. (1986a).

An uncertainty analysis (see Goitein (1985) and Chapter 8) can establish the confidence limits on the dose distribution. Figure 11.6 shows an example of the computed bounds on the penetration of a beam passing through the base of skull (Urie *et al*., 1986a). One sees in this figure the calculated range uncertainty is greater in the shadow of regions of complex heterogeneities, just as one would expect.

To cover the CTV to full dose at a given confidence level (the price being that distal normal tissues receive a greater dose than desired)

one would have to require that the lower bound 90% isodose surface hug the "target volume." If organ motion had been incorporated into the CTV, then the relevant target volume would be the CTV. However, this is usually not, and according to the ICRU definitions should not be, the case. Then, the relevant target volume is the ITV which is defined as enlarging the CTV to account for organ motion within the patient (see Chapter 3).

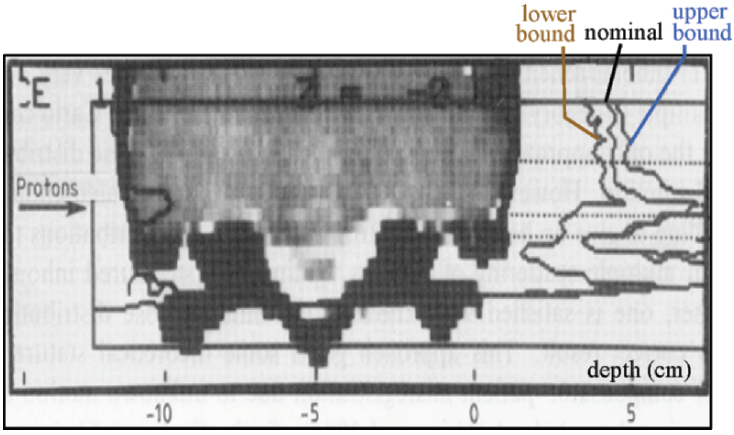


Figure 11.6. Uncertainty analysis for a proton beam traversing a water-filled human skull. The computed upper and lower-bound 90% isodose curves bracket the nominal dose curve – which is the curve which would be estimated in the absence of an uncertainty analysis. Reproduced with permission from Urie *et al*. (1986a).

The Design of Apertures and Compensators

In clinical applications, a proton beam needs to be "shaped" both laterally and in depth. The former is achieved using one or more apertures and/or blocks which intercept protons so that a negligible dose is delivered in their shadow. What little dose there is comes from neutrons produced in the aperture. The latter is achieved using a so-called *compensator* which, in older terminology, was called a "compensating bolus."

Both apertures and compensators can either be physical objects, or they can be virtual – implemented in the latter case by restrictions on the allowed pencil beams in a scanned beam. The principles of their design are rather similar in either case, even though their implementation is completely different. (This is why I have been content to focus on scattered broad beams in much of the preceding discussion; the same principles hold also for scanned beams.)

Apertures

The design of apertures has already been discussed in Chapter 10. It is largely a matter of trivial geometry. For scattered broad beams there are only two wrinkles: there can be aperture-edge effects which give rise to mainly superficial dose perturbations; and neutron production in apertures and/or blocks, while quite small, is not entirely negligible. Both have been discussed in Chapter 10.

Scanned beams do not have either disadvantage. The irradiated volume is entirely defined by the pattern of pencil beams which are applied and there is no material to produce either edge-effects or neutrons. One might think that one would "turn on" pencil beams which are headed towards the target volume, and turn all others "off." One must remember however that scattering effects will deplete the dose of the pencil beams lying at or near the geometric edge of the beam, and so one must add pencil beams around the periphery of the beam in order to ensure that there is no dose deficit to the target volume at its edges. This margin is needed, just as for photons, even if the target volume is the planning target volume (PTV), in order to compensate for dose fall-off in the penumbral region of the beams.

Scanned beams do, however, have one characteristic which must be taken into account. In most implementations, there are two sweeping magnets, spatially offset from one another as shown in Figure 10.21 of Chapter 10. This means that there are two spatially offset virtual sources of the beam – one for each direction of scanning. Thus, the beam's-eye view is more complicated than a simple perspective projection, complicating both the computation of BEV images in the treatment planning system and the design of the aperture.

Compensators

A compensator is a device, be it real or virtual, for making range modifications in the beam that reaches the patient so that wherever a greater beam penetration is required in the patient the compensation is less, and where less penetration is desired, the compensation is more. Figure 11.7a shows schematically the consequence of making no compensation for the bone sliver – one would have a cold region in the target volume. Figure 11.7b shows the consequence of making an "exact" compensation – that is, when the compensator is modified only where it geometrically shadows the inhomogeneity. The dose distribution in Figure 11.7b looks satisfactory (although it ignores scattering effects which would, in fact, prevent perfect compensation). However, if there were a small mis-registration between the

compensator and the patient there would be beam undershoot, and hence reduced dose, in the target volume and beam overshoot in downstream normal tissues, as shown in Figure 11.7c.

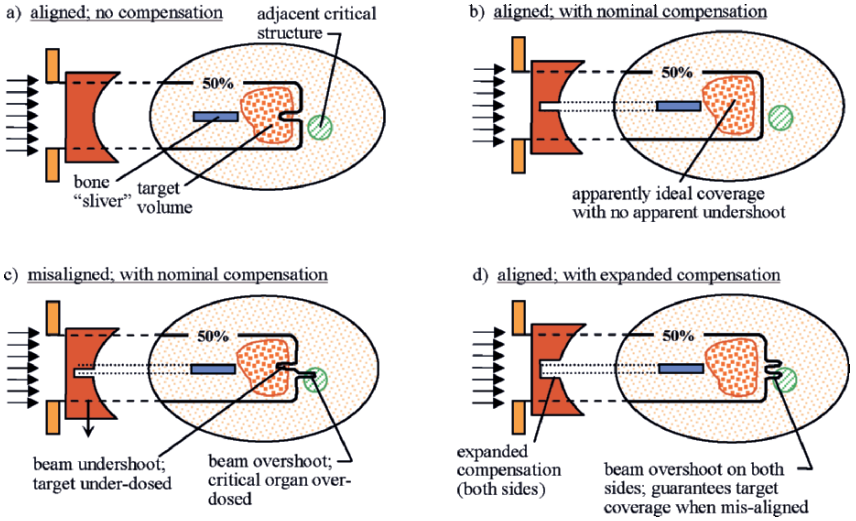


Figure 11.7. Schematic representation of the design of a compensator (see text).

Figure 11.7d shows a possible solution – designed to avoid target underdose at the expense of delivering dose to a larger volume of downstream normal tissues. The modified part of the compensator is widened so that, even if it is slightly mis-registered, the target volume will nevertheless receive full dose. The full prescription for this approach, termed *smearing* of the compensator, is given in Urie *et al.* (1984). Compensator smearing allows not only for mis-registration caused, for example, by patient or organ motion, but also for the blurring effects of multiple scattering.

One useful technique, which is however little used, would be to introduce additional smearing through degradation of the beam *directionality*. Usually, in order to achieve sharp penumbras, the angular confusion of the beam is kept as small as possible. However, as simple inspection of Figure 11.7 would suggest, a spread in beam directions – for example, by delivering a few beams separated in angle from one another by a few degrees – can smear out the dose perturbations beyond an inhomogeneity. In general this strategy will lead to the delivery of a smaller dose perturbation over a larger volume.

The compensation technique described here is relatively crude. Better strategies and computational tools are needed to account for inhomogeneities and for the effects of mis-registration due to motion or changes in anatomy. New strategies are particularly needed for IMPT since its score functions have to be evaluated by a computer, and present algorithms do not reproduce the intelligence of experienced planners who, for example, choose beam directions which avoid "bad" approaches so far as compensator design is concerned.

HU to water-equivalent density conversion

To compute and compensate for the effects of inhomogeneities on proton beams, one needs a quantitative "map" of the patient's tissues. It is more than a coincidence that the growth of proton beam therapy occurred just as computed tomography (CT) became available, for CT provides just such a map. As they are derived from X-ray transmission measurements, CT scan data are in units of relative X-ray absorption coefficients (Hounsfield Units, abbreviated HU). However, detailed measurements have shown that, in analogy with the transformation needed for photon therapy (e.g., Figure 3.5 in Chapter 3), quite satisfactory conversions to proton stopping powers relative to water can be derived from the CT data. Such a conversion is shown in Figure 11.8. The spatial resolution needed for such maps is set by the scale of multiple scattering which is of the order of a few millimeters; happily, the resolution of CT data is reasonably well matched to this.

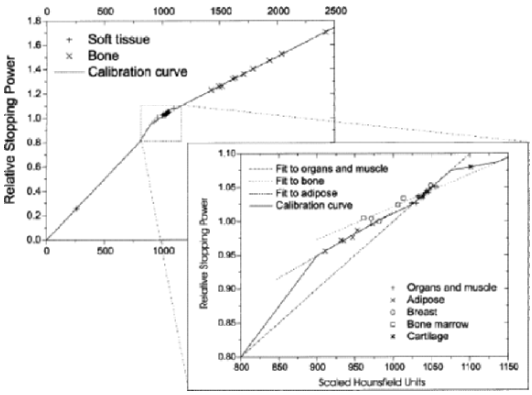


Figure 11.8. Transformation between Hounsfield units and water-equivalent density (relative stopping power) for protons. Reproduced with permission from Schaffner and Pedroni (1998).

In proton beam therapy it is common to refer to the *water-equivalent density* of a material or tissue. This is the density of a fictitious compound with the same chemical composition as water, one

centimeter thickness of which would produce the same range change in a therapeutic proton beam as when it traversed 1 cm of the material in question. The water-equivalent density is luckily quite insensitive to the proton energy.

DOSE COMPUTATION

As with photons, I will not say much about the computation of the dose distributions of proton beams. The estimation of dose has been approached in three ways, in order of increasing accuracy:

broad beam algorithms which compute beam penetration based on modifications of measurements made in homogenous water phantoms, using calculations of the integrated water-equivalent densities along straight lines within the actual patient;

pencil beam algorithms which compute the dose as a super-position of pencil beams (Hong *et al*., 1996). These algorithms can take into account some aspects of differential scattering effects in laterally inhomogeneous materials; and

calculation using Monte Carlo techniques in which programs such as GEANT have been used (Paganetti, 2006; Paganetti *et al*., 2005). Monte Carlo algorithms with more limited physics have also been developed in order to reduce computation times to practicable levels (Tourovsky *et al*., 2005).

RELATIVE BIOLOGICAL EFFECTIVENESS (RBE) OF PROTONS

The biological impact of a given dose of radiation depends on a number of factors, one of which is the microdosimetric pattern of energy deposition. The linear energy transfer (LET) of a proton beam was briefly alluded to in Chapter 10. It is the amount of energy lost per unit distance. The LET varies substantially with proton velocity, and hence position in depth within the patient – as indicated in equation (10.1) and Figure 10.6 of Chapter 10. Coulomb interactions are not, however, the only contributor to LET; nuclear interactions produce charged ion fragments which deliver large doses over small distances – i.e., have large LETs. At any point within a proton beam there will be a spectrum of LETs, and it is thought that the biological effect at a given point within a radiation beam is approximately (but not entirely) related to the average LET at that point.

To account for the biological impact of dose deposition (e.g., the amount of cell inactivation caused by a given delivered dose), the

so-called *relative biological effectiveness*, abbreviated as RBE, is used. The RBE of a given radiation is defined as he ratio of the dose of the reference radiation beam (e.g., photons) to that of the test beam (e.g., protons) required to produce a defined biological response" assuming that all other details of dose delivery – e.g., number of fractions and inter-fraction interval – are the same for both radiations (ICRU78, 2007).

One can then define the RBE-weighted dose as the physical proton dose multiplied by the RBE. This is the dose of photons in the therapeutic energy range which would produce the same effect as the protons, given identical fractionation schemes and end-points. As the RBE is a ratio and therefore unit-less, the unit of the radiobiologically-weighted dose is the same as that of the physical dose, namely Gy. To distinguish between physical and RBE-weighted dose one writes for the former, for example, "a dose of 70 Gy was delivered," and for the latter "a dose of 77 Gy (RBE) was delivered." The addition of the qualifying "(RBE)" indicates that one is giving a statement of the RBE-weighted dose (ICRU78, 2007).[4]

Having said all this, you might imagine that the biological, as opposed to physical, effect of the protons forming, say, a spread-out Bragg peak would vary greatly in depth, since the LET varies in depth. This is not the case, however. It turns out that RBE is relatively constant and near-unity at LETs less than about 200 MeV per $g \cdot cm^{-2}$ and then rises to a value of 3 or more at higher LETs. The average LET of a spread-out proton beam falls almost entirely within the lower range of LETs with only a small component of high LET due to nuclear interactions and to just-stopping protons. Thus, based on the available *in vivo* and *in vitro* laboratory data and on clinical experience, *ICRU78 (2007) recommends using a constant value of 1.10 (relative to ^{60}Co radiation) for the RBE everywhere within a SOBP, including within the entrance plateau.* This has the happy consequence that the dose distributions of proton beams have the same form no matter whether physical or RBE-weighted dose is involved. Only the absolute doses are different, by a constant factor of 1.10.

---

[4] RBE-weighted dose was previously call Cobalt-Gray equivalent dose, and the unit written as CGE. However, this is not an approved SI unit and its use is therefore frowned upon.

I cannot, however, leave things there. It was recognized in ICRU78 (2007) that the above is a simplification and that there are several sources of generally small variations of the RBE within a proton beam around the value of 1.10. Figure 11.9 summarizes schematically where these differences may lie.

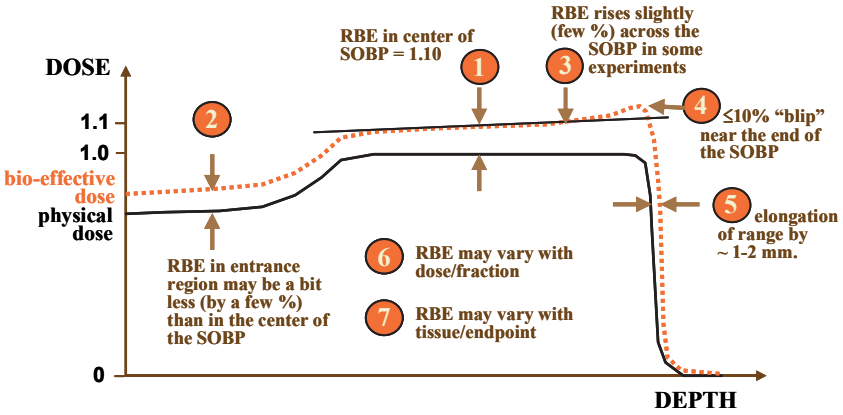


Figure 11.9. Schematic diagram suggesting the ways in which proton RBE may vary from the fixed value of 1.10 recommended by ICRU78 (2007). The "blip at the distal end of the SOBP is, in reality, not a separate phenomenon, but a region in which the average proton energy becomes increasingly low as the depth gets larger – so that the LET (and, hence, RBE) becomes therefore increasingly higher with increasing depth.

Of all these effects, the greatest uncertainty is in item 6, the variation of RBE with fraction size. Both on theoretical grounds, and from extrapolation of *in vitro* experiments, one would expect the RBE to rise as fraction size is reduced. However, neither the scant *in vivo* data nor clinical experience seem to exhibit such a behavior at therapeutic dose levels (Paganetti *et al*., 2002).

In the future one hopes and expects that the various effects identified in Figure 11.9 will be better understood and quantified. If so, and if international agreement can be reached as to how to estimate them, a more nuanced RBE estimate will be possible.

PLANNING PROTON BEAM TREATMENTS: WHAT'S DIFFERENT?

At this point I have devoted many pages to making the point that proton dose distributions are different from photon dose distributions because the physics of the interactions is different. And that, with

protons, one can control the deposition of dose along the direction of the beam as well as laterally. So, the question is, when it comes to planning proton beam therapy, what's all the fuss about? Why isn't

Table 11.1. List of the steps in the planning process which differ in proton beam planning as compared with photon beam planning.

step		*protons vs. photons*
1	Evaluate the patient using all relevant diagnostic tools, and decide whether to employ radiation therapy.	~same (but protons may affect choice of modality)
2	Obtain and inter-register imaging studies with the patient lying in the position to be used for therapy.	same
3	Delineate on the planning CT the target volumes (GTV, CTV and PTV) and normal tissues.	~same (but PTV has different interpretation)
4	Establish the planning aims for the treatment.	same
5	Design one or more sets of beams, together with their weights, each of which fulfills, to the extent possible, the requirements of the planning aims.	different
6	Evaluate these plan(s) and either select one of them for use or revise the planning aims and return to step 5.	same
7	Finalize the prescription.	same
8	Simulate the selected plan to ensure it is deliverable.	same
9	Deliver the treatment, and verify that the delivery is correct.	~same (but QA harder)
10	Re-evaluate the patient during the course of treatment and, if necessary, return to step 5, or even 2, to re-plan the remainder of the treatment.	same
11	Document and archive the final treatment plan.	same
12	Review the treatment plan at the time of patient follow-up or possible recurrence.	same

the planning process just the same as for photons, except that proton beams have a different depth–dose distribution?

In a certain sense, there is some truth to that last claim. However, the devil is – as always – in the details. Proton beam therapy has acquired a reputation of being harder to plan than photon therapy, and I want here to describe what the differences seem to be. Let us start with the list of planning tasks laid out in Chapter 6 in the context of photon beam therapy. Which tasks are different when protons are used? Table 11.1 reproduces these steps and identifies those which differ.[5]

One can see at once that the steps that are the same far outweigh those which are different; There is only one, admittedly large, step for which the planning of proton therapy is substantially different from planning photon therapy.

One can say, in a nutshell, that the differences arise from three causes, the first two of which are related:

- the finite penetration of protons in matter;
- the sensitivity of protons to the traversed materials both in terms of their penetration and their lateral characteristics;
- the different beam formation techniques used in proton therapy.

The differences manifest themselves in a number of ways. The following sections offer a brief description of the differences – many of which, as I have touched on them already, are mentioned without additional elaboration.[6]

DIFFERENCES IN PLANNING: STEP 1 – CHOICE OF MODALITY

Step 1 of Table 11.1 involves the decision as to whether radiation therapy is appropriate for the patient. One aspect of this question relates to the choice of radiation modality. It may even be that a

---

[5] I would like to acknowledge the input, which was given in the context of preparing ICRU78 (2007), of a number of colleagues in addressing this question, namely: J. Adams, M.Moyers, P. Petti, S. Rosenthal, B. Schaffner, N. Schreuder, and L. Verhey.

[6] The issue of the interpretation of the PTV, mentioned in step 3, is addressed in the discussion of step 5 issues.

given modality (protons in our case) may make radiation a good choice, whereas, for other modalities, radiation may not be the preferred choice. In deciding whether protons are likely to be helpful, two important considerations need emphasis. These are:

Large targets

Protons have acquired a reputation for being particularly useful for small targets. In my opinion this is an incorrect perception; protons are probably more useful for the treatment of large targets than small ones. This is because of dose–volume effects in the tissues outside the target volume as discussed in Chapter 5. The larger the target, the smaller the remaining volume of normal tissue and, therefore, the greater the need to spare it. Hence, the larger the target, the greater the clinical advantage of the dose–sparing properties of protons is likely to be. Paradoxically, the excellent results in the treatment of ocular melanomas with protons are a good example of this principle. For, although the treated volume is physically small, it can be quite a large fraction (up to at least one-third) of the relevant body compartment – namely, the eye.

Complex geometry

Protons are good at solving patient-specific problems in which the geometric relationships between the tumor and adjacent OARs pose difficulties and normal tissue avoidance is important. The excellent results in the treatment of base-of-skull sarcomas with protons are a good example of this principle, for these tumors often wrap around, or are very close to, sensitive normal tissues such as the brain stem and optic chiasm.

DIFFERENCES IN PLANNING: STEP 5 – DESIGN OF BEAMS

The effects of inhomogeneities

I have already devoted considerable space to a discussion of the effects of inhomogeneities and I will not repeat the points already made. However, in addition to the obvious fact that inhomogeneities alter the penetration of protons in their shadow, and that one needs to alter the beam by designing real or virtual compensation to take account of this, there are a few additional points worth making:

Uncertainty One cannot predict all the effects of inhomogeneities exactly. Mis-registration of the compensation scheme relative to the patient can arise due to patient and organ localization

localization changes and uncertainties; beam scattering means that perfect compensation is impossible; measurements of the location and nature of the inhomogeneities are imperfect; and so forth. As a consequence, one is obliged to make a relatively complex analysis of the uncertainties and then design the beams so as to make the clinical consequences of those uncertainties as acceptable as possible. Of course, the same is true of photon beam therapy, but the proton problems are more complex and require more complex solutions since correction in depth as well as laterally is necessary.

Heterogeneous inhomogeneities With one exception, I have treated inhomogeneities as though they were internally homogeneous. The exception is the complex situation in the base-of-skull, discussed above. You will recall that the distal part of the proton beam was not just shifted in range, but was smeared out. This was due, presumably, to the many possible paths, each with a somewhat different water-equivalent path length, that the scattering protons can follow. In (Urie *et al*., 1986a) a dramatic degradation of the distal beam was seen in the abdomen – probably mainly due to organ motion. And similar smearing has been observed in the lung (R. Mohan, private communication) and in granulated graphite (S. Vynckier, private communication). The possibility of distal smearing by complex inhomogeneities, which can be up to at least ±2 cm, must be taken into account when designing a proton beam.

Over-penetration in the lung It is usual, in using protons, to add a safety margin in depth (i.e., in energy), as well as laterally, to account for the various uncertainties. This may be, say, sufficient additional energy so as to provide in near-unit density materials such as muscle or brain an extra 0.5 to 1.0 cm of penetration beyond the target volume. However, lung has a low water-equivalent density (let us use 0.2 $g \cdot cm^{-3}$ for the purpose of illustration) which has the consequence that the same increase in proton beam energy would lead to a factor of about 5 greater penetration – that is, to a 2.5 to 5 cm overshoot in lung. Moreover, organ motion can further increase this margin, due to the possibility of tissues moving in and out of the beam path with the respiratory- and cardiac-induced motion. Thus, after allowing for the inevitable uncertainties, one may find that one has to treat an undesirably large volume of lung, and/or that critical structures distal to the lung may receive unwanted dose. One must strive particularly hard to reduce the uncertainties when protons must traverse lung, so as to be able to limit the safety margin to a small physical distance. Clearly

respiration gating is a "must" in proton treatments involving the thorax, and the same holds in the abdomen where diaphragmatic motion similarly affects the location of organs.

Metallic or other high-Z implants Metallic implants are quite common, either as prostheses or in the form of surgical aids such as clips, in patients when they come for proton beam therapy. These implants cause problems for two reasons. First, they cause severe artifacts in the CT scans which can cause errors in computing path lengths within the patient and, therefore, can result in erroneous compensator designs. And second, their water-equivalent thickness is usually inaccurately gauged, both because of lack of transmitted photons in the CT scanning, and because the conversion of Hounsfield number to water-equivalent density breaks down. The best solution – nowhere implemented at the time of writing – seems to be to use a CT scanner which employs a megavoltage source of radiation. This would largely solve both of the above problems.

Compensation for inhomogeneities

The approach to designing a beam compensator for proton beam therapy has already been described. It requires allowance for registration errors and other uncertainties, and allowance for the fact that scattering of the protons in the patient and in the upstream material prevents perfect compensation.

Choosing "good" beam directions Perhaps the most fruitful approach is the avoidance of beam directions for which inhomogeneities pose the greatest difficulties. This is something that experienced planners do automatically, but it is only just now receiving serious attention in the case of computer-driven planning – under the rubric of "robust planning." In IMRT, especially when using photons, there is a tendency to pick a set of equally spaced beam directions without attention to patient-specific geometry. This does not seem like a good idea with protons. Beams which go near-parallel to boundaries between materials of very different density, including the air/tissue interface at the skin surface, or which pass lengthways through a complex inhomogeneity such as the petrous ridge, should be avoided.

Angular feathering As a last resort, one can and should use a few beams, closely separated in angle by a few degrees, instead of a single beam when one cannot avoid passing near-parallel to an inhomogeneity (Goitein, 1977).

Feathering in depth When two proton beams abut at the end of range of one or both beams, one should employ depth feathering – that is the use of a few beams, each closely separated in range by several millimeters, instead of a single beam. This is done in beam-patching which is described below.

Beam delivery techniques

A treatment planning program needs to simulate the full variety of beam delivery techniques available for each treatment machine. For protons, these are quite varied – and quite different from those of X-ray beams. Proton beams can be delivered by scattering, wobbling, or scanning techniques. Then, too, proton beams can be shaped both laterally and distally. The former by apertures and blocks, the latter by distal shaping using a compensator or, when using IMPT without patient-specific hardware, by 3D shaping achieved by adjusting the intensities of the scanned pencil beams.

One particular point of difference, already discussed in Chapter 10, is that the dose distribution is sensitive to whatever upstream material is in the beam. As a result, the planner must pay greater attention to locating patient-specific devices at an optimal distance from the patient. One wants them close, to minimize the effects of scattering in the compensator and to have as small a penumbra as possible, but not too close, because of edge-scattering in the aperture if there is one.

The planning target volume (PTV)

Being an alert reader, you have probably noticed that I often use the term "target volume" without being specific as to whether I am referring to the CTV or the PTV. In part, this is because the PTV is difficult to define in the case of protons and it may be of less use in designing a beam than it is in the case of photon treatments.

In the case of photon beams, the PTV is primarily used to set the lateral margins of the field in order to compensate for motion and setup uncertainties – and, even then, its use has to be supplemented by knowledge of the characteristics of the beam penumbra. A single PTV can set the margin no matter the direction of the beam.

In the case of proton beams, *both lateral and distal margins are needed* – the former to set the lateral margins of the field in order to compensate for motion and setup uncertainties, the latter to set the proximal and especially distal margins of the treated volume in order

to compensate for uncertainties in the penetration of the protons. These two margins are, generally, quite different from each other. As a result, one cannot design a single PTV which could be used to set both the lateral and distal margins for all beam directions. One would have to design a separate PTV for each possible beam direction. This is not presently done as it would be computationally too time consuming. Consequently, proton beams are often designed directly from the CTV, as seen in the beam's-eye view, with the design of both lateral and distal margins being built into the algorithms which set the beam shape. It has been suggested (ICRU78, 2007) that the CTV might be used for beam design and the PTV, encompassing only lateral margins, be used purely for the purpose of dose reporting so as to facilitate uniform practice in both proton and photon radiotherapy.

Design of single beams

Although uniform-beam treatments are usually planned manually, it is interesting to note that the construction of apertures and compensators are inherently inverse processes. That is, from a statement of the problem "Cover the target volume with a specified high dose." the aperture and compensator are designed automatically in a generally one-pass process.

In what has been said so far, I have emphasized the goal of compensation design as being to ensure full target volume coverage. Of course, the opposite could also be the case. One might want to avoid irradiation of a specific organ by more than some specified dose or, better, dose–volume condition, even at the expense of underdosing the tumor. And, even more likely, one might want to strike a balance between these extremes. This then leads to a consideration of optimization of proton beam therapy – about which I will shortly make a few remarks. The point here, however, is that the goal of compensation is itself a variable of the planning process and the compensation scheme should be designed so as to achieve the clinical objectives, as stated in the planning aims.

The distal fall-off of a proton beam of, say, 150 MeV, is inherently, on the basis of the physics of protons impinging on a bucket of water as discussed in Chapter 10, steeper than the lateral fall-off. However, this is theory; practice is different. The uncertainties in the distal penetration of protons in the complex situation of patient treatments, usually result in the *effective* distal fall-off being quite a bit greater than the lateral fall-off. This is the reason that one often prefers to take advantage of the lateral edge of a collimated beam in protecting a

sensitive organ lying close to the target volume, rather than the distal edge.

It goes without saying that the algorithm for dose computation is quite different for protons than for photons. However, this is largely transparent to the planner. Given the complexities of the effects of inhomogeneities, however, the urge to use a Monte Carlo algorithm is rather greater with protons.

The RBE of protons, as discussed above, is not a problem in practice, given the fortuitously simple recommendation that a constant value of 1.10 be used everywhere. It is, of course, essential that, wherever absolute doses are shown, it is made absolutely clear as to whether physical or RBE-weighted doses are at stake. In general, ICRU78 (2007) recommends using RBE-weighted doses – identified by including the qualifier "(RBE)" after the dose statement – pretty much everywhere.

Nevertheless, a planner should be aware that, when bringing the distal end of a proton beam up close to a sensitive organ: (a) the effective dose may extend from 1 to 2 mm beyond the physical dose: and (b) that there may be an elevation of the effective dose in the last several millimeters of range – that is, the "blip" featured in Figure 11.9.

Design of plan(s)

Once the differences discussed so far are taken into account, the design of a plan of treatment (a set of beams and beam weights) proceeds very much as for photons. One has manual and automatic approaches in both cases. For manual planning, the issue is largely the choice of the number, direction, and shape – including for protons the shape in depth – of the beams and, of course, of their weights. As with photon beams, as discussed in Chapter 8, the use of non-coplanar beams (beams whose central axes do not lie in a single plane) can be very advantageous. All the tools needed to create a set of beams, to assess the resulting dose distribution, and to compare rival plans are the same.

IMPT is, in principle, just as easy (or hard) to design and implement as *intensity-modulated X-ray therapy* (IMXT). The computational task is made greater, but not really more complicated, by the greater number of degrees of freedom involved in IMPT. In essence, one has a set of pencil beams which cover the area of the projected target volume, as is also the case in IMXT, but then, for each location in the

field, one has a set of pencil beams of varying penetration. Typically, some 20 beam energies might be involved, irradiating discreet "layers" spaced *circa* 0.5 g·cm^{-2} apart. This could potentially make the task of computation one and a half orders of magnitude larger and, hence, longer. However, there are some tricks available to speed things up, such as ways to reduce the number of pencil beams which are tried out.

As previously mentioned, in proton beam therapy the choice of beam angles is more important than for photons and, therefore, unlike IMXT, the use of equally spaced angles is unusual. Generally speaking, for both IMPT and for uniform beam proton therapy, the number of beams needed to produce a satisfactory plan is typically less than the number needed for X-rays (Rutz and Lomax, 2005). The use in IMXT of several (e.g., 5, 7, or 9) beams, equally spaced in angle and non-collinear, tends to result in spreading the dose outside the target volume over pretty much the full 360° available – and favors coplanar beam arrangements. IMPT, on the other hand, can use rather few beams and, thus, can distribute dose over a smaller volume of the normal tissues. This brings up the question, already raised in Chapters 8 and 9, of whether it is better to spread out the energy outside the target volume, or to concentrate it in a smaller volume at a higher dose level. I have emphasized that we don't really know the answer to this question – but, at least, one has a choice with IMPT, whereas IMXT is likely to result in near-360° dose spreading.

Immobilization, localization, and verification

Proton beam therapy has always emphasized the need for spatial as well as dosimetric accuracy in treatment delivery, and hence planning. This has arisen from the need to reduce compensator registration errors, and to have the irradiated volume be as close as possible to the target volume. As a consequence, traditionally greater attention has been paid to patient immobilization, localization and verification than in the case of photon therapy. However, this situation has changed in recent years and the advent of first 3DCRT and then IMXT has brought photon therapy much more in line with proton therapy practice.

Uncertainty analysis

I fear that here I am beating a dead horse, but for completeness I have to reiterate that, while uncertainty analysis is essential for all forms of

radiation therapy, it is particularly important in proton beam therapy for all the reasons cited previously.

DIFFERENCES IN PLANNING: STEP 9 – QUALITY ASSURANCE

In Chapter 12, I emphasize the important role of quality assurance (QA) in radiation therapy in general. QA is no more or less important for proton beam therapy than for other radiation modalities. However, there are a few differences.

First, the process of verifying the dose delivered to, say, a bucket of water, is more intensive because there are more variables; one must check with spatially fine resolution, the fidelity of the dose in depth as well as laterally. This means that one needs to be able to measure the dose in 3D. Moreover, there is a severe time constraint in the case of scanned or wobbled beams, namely: it takes of the order of a minute to deliver a single beam. With only a set of separate 2D (or, even more so, 1D or 0D) active detectors, one needs to deliver the same beam many times in order to build up a 3D dose distribution from the measurements. This is very time-consuming and makes routine QA quite difficult and potentially time consuming. On the other hand, there is currently, unfortunately, no good practical 3D dose measuring device. One is badly needed.

So one is left at the moment with either a sequence of 2D measurements, or the use of a closely paced array of ionization chambers which, for practical reasons, cannot be numerous enough to provide the spatial resolution one would like.

A scintillation screen viewed by a digital camera provides a 2D dose distribution measurement with good spatial resolution (Boon *et al*., 2000). An alternative form of 2D detector is film. This can either be radiographic film in contact with a scintillating screen, or Gaf-Chromic film which has the advantage of being self-developing. With a 2D detector one would ideally like to make, let us say, measurements at 30 different depths. Since each exposure takes about a minute (due to the length of time needed to deliver a beam), it is very is very time-consuming to acquire a 3D dose distribution.

There is interest in treatment verification using PET imaging. Protons induce radioactivity in the irradiated material as a result of non-elastic collisions with atomic nuclei (see Chapter 10), and some of that activity is due to the formation of positron emitting radioisotopes. The concentration of positron emitters at a point is initially related to the dose deposited at that point. PET scanners can detect the distribution of induced activity and hence provide a confirmation of

the accuracy of the dose distribution delivered to the patient. There are three caveats. First, activity is carried away from the site of creation by blood flow so that, by the time of imaging, the PET image may be washed out and not accurately reflect the dose distribution. Second, the amount of activation depends on the proton energy. In particular, there is no activation at the very end of range of the protons – although this problem can be mitigated by modeling. And, third, the spatial resolution of PET scans (of the order of a few millimeters) is not quite good enough to detect dose perturbations within small volumes which, as we have seen in Figure 11.4, may occur in distances as small as a millimeter.

Finally, proton radiography, although it has so far not been used in practice, has the potential to verify proton dose algorithms and even, perhaps, be used to modify a patient's plan on a day-to-day basis (Schneider and Pedroni, 1995). Proton radiography involves measuring the residual energy of protons exiting the patient (or a phantom) using a position-sensitive range telescope of some kind. Since proton beams used for treatment normally do not exit the patient, one has to add energy to the beam for the purpose of making (low dose) transmission measurements. One can then determine the water-equivalent path length through the patient. This is not the same as that to the distal target volume surface, but good agreement with the calculated exit energies would build substantial confidence in the treatment delivery.

DOSE DISTRIBUTIONS ACHIEVABLE WITH PROTONS

Scattered beams

We have seen that even a single proton beam (e.g., Figure 10.1, left panel), in contradistinction to a single photon beam (e.g., Figure 10.1, right panel), can provide an acceptable treatment. However, just as for photons, the use of multiple cross-firing proton beams focused on the target volume reduces the dose delivered outside the target volume – at the cost, of course, of spreading the inevitable energy over a larger volume. Figure 11.10 schematically illustrates the difference between protons and photons for one-, two- and four-field approaches designed to deliver 60 Gy (RBE) to the target volume. (The top row of this figure is identical to the sketches in Figure 1.3 of Chapter 1.) No matter how many fields are used, the proton dose outside the target volume is substantially lower than the photon dose – with the exception of the proton's lack of skin-sparing

at the beam entrance. While the numbers in Figure 11.10 are very approximate, detailed treatment planning comparisons fully bear this point out. The integral dose delivered outside the target volume is typically between two or more times lower for protons than for photons (Lomax *et al.*, 1999).

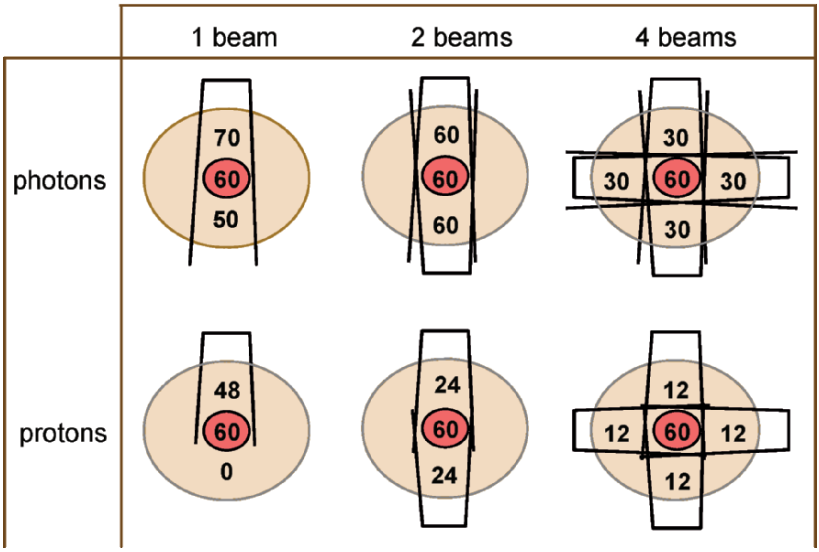


Figure 11.10. Schematic diagram comparing 1-, 2- and 4-beam treatments of photons (*top*) and protons (*bottom*). In both cases, the use of multiple fields reduces the dose outside the target volume while spreading it out. The integral dose outside the target volume due to protons is, in this very crude example, 2.5 times less than from photons.

A practical example of the dose distribution of a treatment of a base of skull sarcoma, using scattered protons, is shown in Figure 11.11. The white diamonds show the intersection of the target volume (CTV) with the coronal plane. The colored lines show the field outlines as they intersect the plane. This looks like an attractive dose

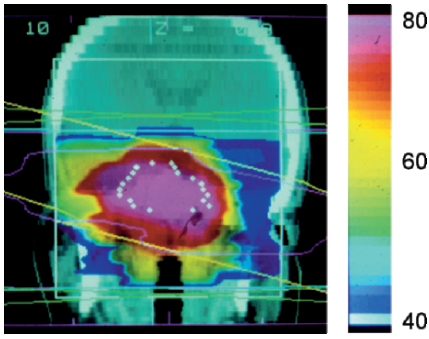


Figure 11.11. Coronal section of a proton treatment of a base-of-skull sarcoma employing six non-coplanar beams (doses in Gy (RBE)).

distribution but, on closer inspection, the question occurs to one as to why the high dose region extends so far outside the target volume. There are two reasons for this which reinforce one another. First, as I have already discussed *in extenso*, a safety margin has been left and, because many of the beams are going through very complex inhomogeneities in the base of skull, the distal margins in particular can be quite substantial. Second, it is often the case that a neighboring section has a larger target area. Because of the need to provide safety margins in all directions, the dose needed to adequately cover the neighboring section "spills over" into the section one is looking at. A particular example of this arises when one is using a scattered proton beam to treat a pear-shaped volume from the top of the pear. In order to cover the cross section of the target volume at depth, a larger-than-desirable field covers the smaller upper sections of the pear. It is very important in judging dose distributions to inspect the dose in many sections throughout the treated volume, and not just in one or a few selected sections.

The point I am making here is a very fundamental one. Too frequently one sees charged particle plans in which the high dose volume tightly hugs the target volume. Such plans are testimony to a lack of understanding of the importance of uncertainty estimation and too much faith in the pretty colored pictures which can be painted by the computer. I am convinced that the good local control that has been achieved at the Massachusetts General Hospital in treating base-of-skull sarcomas has owed a lot to the conservative treatment margins employed.

Field patching

One not infrequently encounters the situation in which a horse-shoe shaped target volume is wrapped around an organ to which one does not want to deliver the full prescription dose. This problem can be solved in one of several ways. IMPT is one, as discussed below, and the use of a single proton field tailored distally to stay off the central region of the horse shoe (e.g., the anterior field sketched in Figure 11.12a) is another. However, this second approach can be highly problematic if the tissue densities proximal to the organ to be spared are complex, and it is often avoided for that reason. The beam whose 90% isodose contour is outlined in Figure 11.12a, for example, would never be considered in practice because of the complex inhomogeneities through which it passes. One could not be sure of sparing the brain stem.

To meet this problem a technique of "field patching" has been developed. Figure 11.12 shows an example of this method. In this example, a lateral beam (Figure 11.12b) is combined with a pair of posterior "patch" beams (Figure 11.12c). The composite dose distribution (Figure 11.12c) shows an elevated dose in the junction region. The problem of patching is highlighted when one looks at the uncertainty analysis and sees the potential for either hot or cold spots in the plan (Figures 11.12d and 11.12e). For this reason, the posterior patches are in practice feathered, as described above.

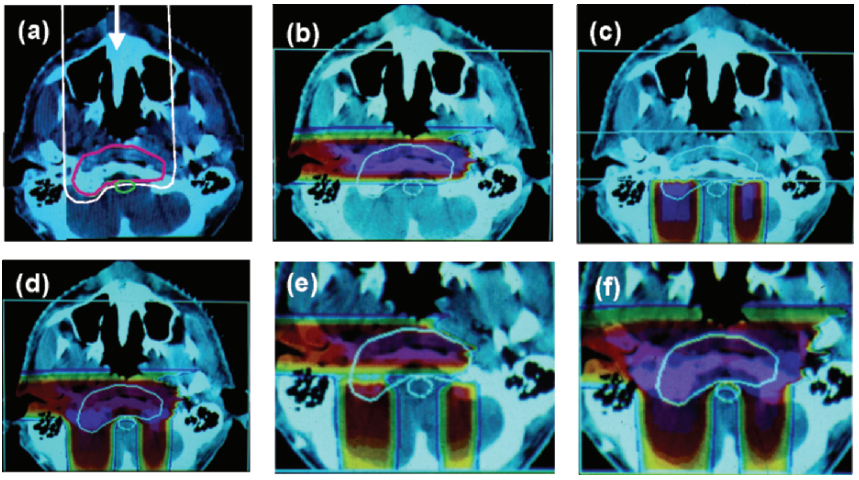


Figure 11.12. Example of patched proton beams used to treat a horse-shoe shaped target volume (CTV): (a) outline of a possible anterior field which, however, would be contraindicated due to inhomogeneities; (b) lateral field; (c) pair of posterior patched fields; (d) composite plan; (e) lower bound dose distribution; and (f) upper bound dose distribution.

Patched fields are, in fact, early examples of intensity-modulated radiation therapy in that each individual field irradiates the target volume non-uniformly.

Intensity-modulated proton therapy

Passive scattering is a mature approach which offers a simple and effective method of delivering proton therapy. For a single field direction, passive scattering can provide excellent conformation of dose to the distal end of the target and good conformation laterally. However, due to the fixed depth-modulation of Bragg peaks across the whole field, it neither can provide conformation of the dose to the

proximal surface of the target volume, nor can it modify the intensity within the target volume.

Beam scanning can be used to deliver any physically possible dose distribution. In particular, it can be used to deliver IMPT. Indeed, scanning was developed with precisely this application in mind. With beam scanning, it is possible to vary the intensity of the pencil beams both across the field and in depth and, therefore, one can "turn off" pencil beams which terminate proximal to the target volume and hence contribute no useful dose to the target. In this way the high dose volume can be tailored to the target volume proximally as well as distally.

Figure 11.13 shows three different dose delivery techniques, all using scanned beams, for both a single beam and for a 3-field plan.

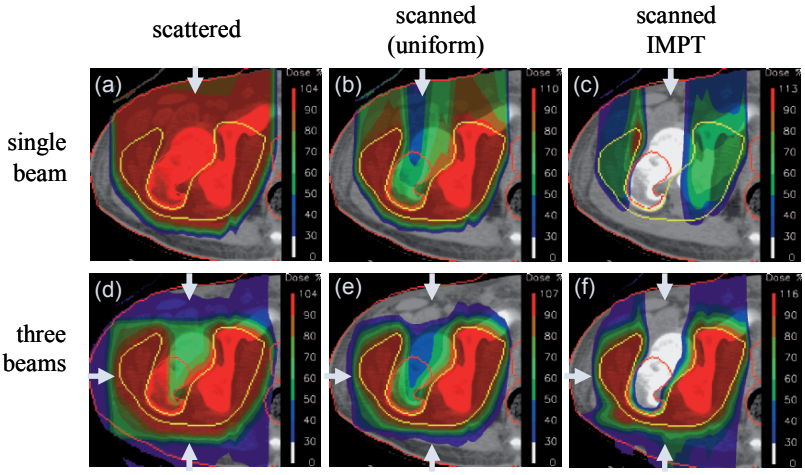


Figure 11.13. Dose distributions for single beam (*top*) and three-beam (*bottom*) proton treatments. *Left column (a and d):* simulated scattered beams; *middle column (b and e):* scanned beams, but delivering uniform dose to the target volume; and *right column (c and f):* IMPT. Figure adapted with permission from Goitein *et al*. (2002)

The three techniques shown are: column 1: mimic of a scattered beam – i.e., a beam which is uniform across the field and has a fixed modulation in depth across the field (mimicking a SOBP); column 2: beams as in (1) except that those pencil beams which terminate proximal to the target volume are turned off, thus shaping the high dose region proximal to the target volume; and

column 3: fully-fledged IMPT. The comparative DVHs for these plans are shown in Figure 6.11 of Chapter 6.

The proximal dose sparing seen in panels (b) and (e) does not provide as great an advantage as it might at first seem, except in the case of very complex target volumes. Goitein and Chen (1982), for example, showed for elliptical target volumes that proximal dose sparing reduced the integral dose by only about 10% or so in proton beam therapy. However, for complexly-shaped invaginated target volumes, savings of integral dose can be more substantial.

A somewhat more complex example of IMPT is shown in Figure 11.14, which is an example of a proton plan using IMPT to treat a nasopharyngeal carcinoma. This figure should be compared with Figure 9.2 in Chapter 9, where an example of IMXT for the same patient was shown.

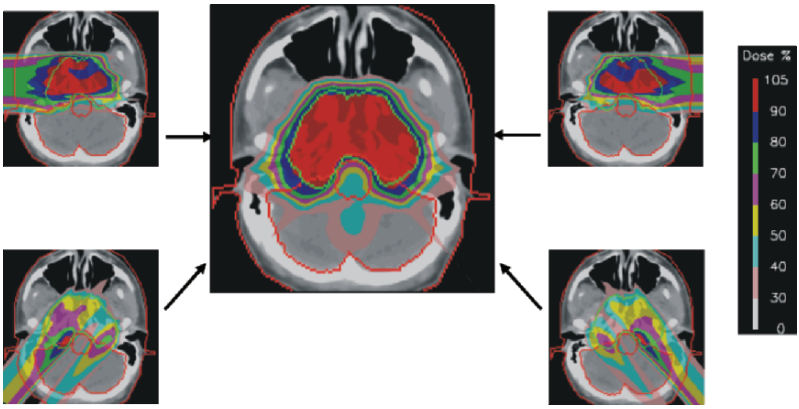


Figure 11.14. Four field IMPT treatment of a nasopharyngeal carcinoma. Figure courtesy of A. Lomax, PSI, CH.

It is noteworthy that an excellent, indeed superior, dose distribution is achieved with only four proton beams – as opposed to the nine photon beams used in the case shown in Figure 9.2. This seems to be generally the case. Fewer proton beam directions are needed to treat a given target as compared with photons. This is presumably due to the added degree of freedom (control of range) that protons enjoy.

As a special case of IMPT, it has been suggested by Deasy *et al.* (1997) that proton treatments be given using only pencil beams which have their Bragg peak located on the distal surface of the target (*distal-edge tracking*). This approach, for centrally situated tumors, is

thought to minimize the total integral dose delivered to the patient, sharpen the lateral fall-off of the resulting dose distribution and can be more quickly calculated and delivered due to the significantly reduced number of Bragg peaks that need to be delivered. However, one would expect that, if this is truly the optimal technique, that optimization programs used for IMPT would converge to it and there would be no need to "guide" the solution towards the distal-edge tracking geometry.

Figure 11.15 presents a side-by-side comparison of photon and proton beam dose distributions.

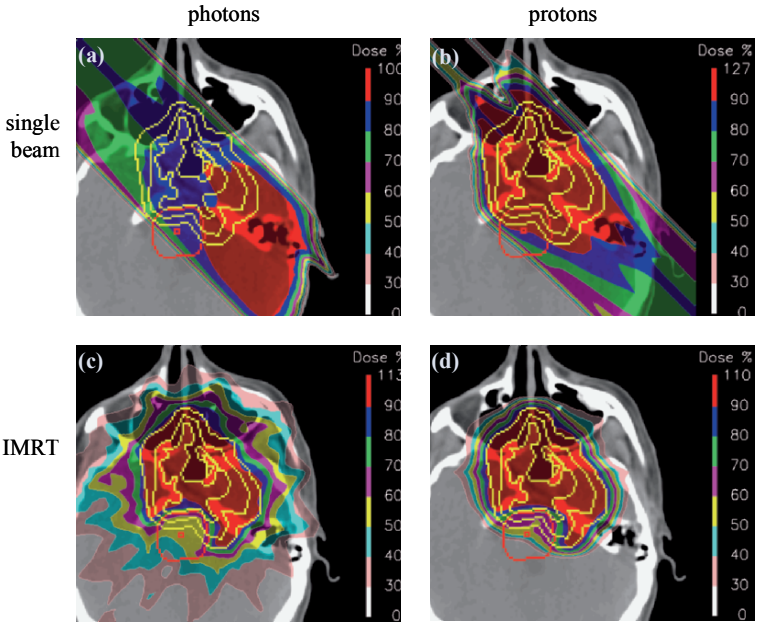


Figure 11.15. Example of a meningioma treated by photons (*left side*) and protons (*right side*). The three volumes are the GTV, CTV and PTV. Panels (a) and (b) show single left posterior oblique fields, and panels (c) and (d) show the full IMRT plan for (c) photons and (d) protons. Figure courtesy of A. Lomax, PSI,

The largest difference, as already emphasized, is that photons deliver a substantial excess "dose bath" outside the target volume. This point is made again, even more explicitly, in Figure 11.16 which shows a section of a plan to treat a large Ewings sarcoma with either IMPT or IMXT. It is hard to imagine anyone wanting voluntarily to receive

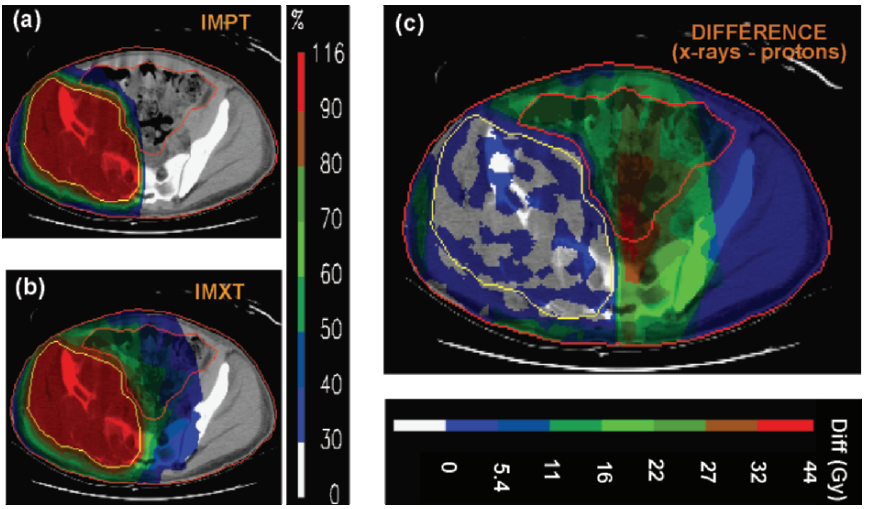


Figure 11.16. Comparison of: (a) IMPT, and (b) IMXT in the treatment of a large Ewings sarcoma. The color scale for panels (a) and (b) is shown to their right, and is in relative dose. Panel (c) shows the dose difference between the two plans in Gy (RBE), assuming a prescription dose of 54 Gy (RBE). Panel (c)'s color scale is directly below it, and is in absolute dose. Figure courtesy of A. Lomax, PSI, CH.

the additional approximately 15 to 30 Gy (RBE) to the intestines which photons deliver in this case, as shown in Figure 11.16c.

Numerous, much more fully documented, comparisons of proton beam with photon beam dose distributions have been published. Glimelius *et al*. (2005) have compiled a list of references to 52 such comparisons. It is fair to say that the vast preponderance of these comparisons confirm that the physical characteristics of protons dictate that the integral dose delivered to a patient will always be reduced in comparison to that delivered by photons. The quantitative reduction by protons of the integral dose delivered outside the target volume by a factor of two or more is largely independent of the delivery method used; whether or not IMRT is used does not substantially reduce the total delivered energy – except in the case of complexly-shaped target volumes where IMRT can have an integral dose advantage over uniform-intensity radiation therapy.

TREATMENT OF OCULAR MELANOMAS

If only because the treatment of ocular melanomas has been one of the largest clinical experiences with proton beam therapy and has been very successful, I will say just a few words about these specialized treatments (Gragoudas *et al*., 2002; Egger *et al*., 2003; Goitein and Miller, 1983).

Figure 11.17 shows several steps in the treatment process: Panel (a) shows a wide-angle fundus view of an ocular melanoma; often ocular melanomas are first discovered upon fundus examination. Panel (b) shows a pathological specimen and, schematically, how at operation when the posterior of the eye is exposed, the eye can be trans-illuminated. Panel (c) shows radio-opaque clips, some 2 mm in diameter, which are sutured to the sclera around the periphery of the tumor. The location of the clips relative to the tumor base can be seen during trans-illumination.

A computer model of the eye is built up as in panel (d); the normal structures are taken from a library of structures of interest, scaled to the dimensions of the eye as measured on A-mode ultrasound. The tumor base is drawn on the surface of the retina as shown in panel (e), based on drawings of the tumor-to-clip relationships made at the time of operation, and on the tumor shape as seen in the fundus picture. The body of the tumor is then added as seen in panel (f), based primarily on ultrasound to measure the tumor's height and shape and on visual examination of the eye.

A direction of gaze is chosen by having the computer interactively move a virtual light source around with the eye following the light until a desirable interrelationship of the target volume and the normal structures, as seen in the beam's-eye view, has been achieved. This process is based on the planner's experience. An aperture is then drawn (panel (g)) with a margin of the order of 2 mm to allow for sub-clinical extension of the tumor, alignment uncertainties, and the 90-50% penumbral width, which is typically about 1 mm. The dose distributions can then be examined in sections through the eye (panel (h)) and as isodose lines drawn on the curved retinal surface, as seen in panel (i) which can be compared with panel (a). Dose–volume histograms (panel (j)) can also be inspected.

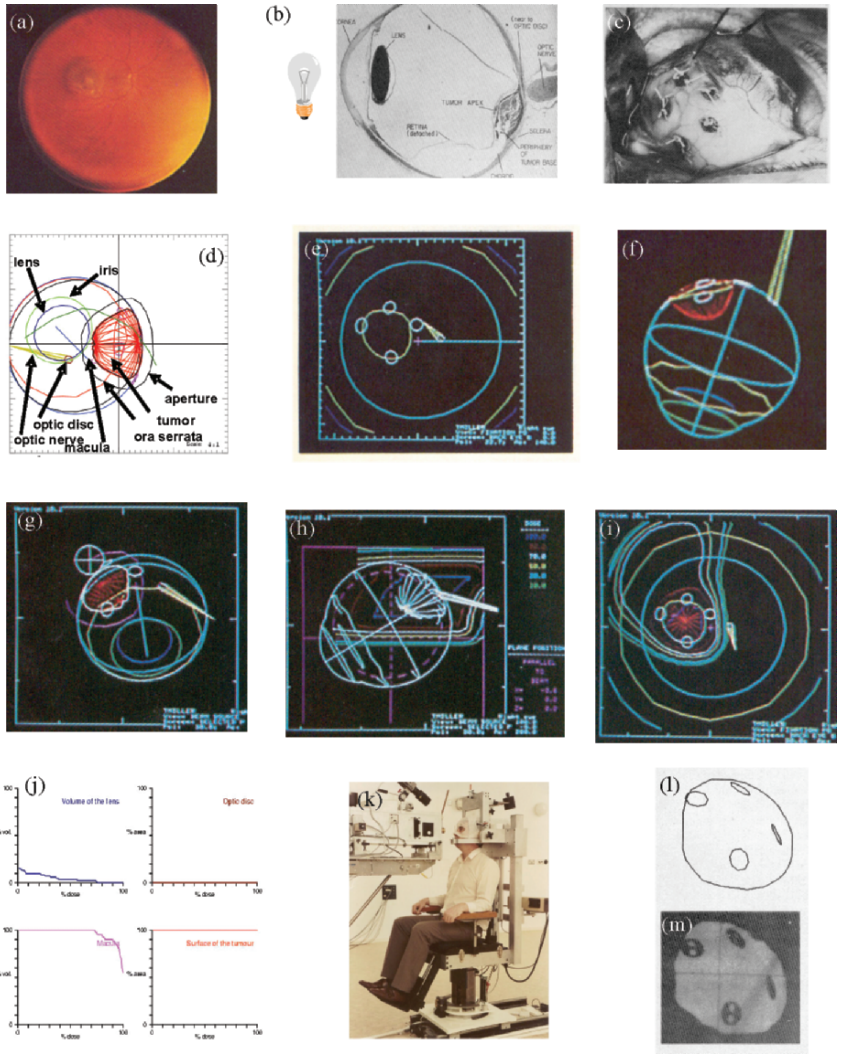


Figure 11.17. Images describing the work-up and treatment of ocular melanomas by protons (see text). Panels (d), (j) and (k) courtesy of J. Vervey and G .Goitein, PSI, CH.

For treatment, the patient is seated in a specially designed chair (panel k) which can move in all three directions of translation and his or her head is immobilized in a mask with bite-block which can be rotated and tilted (pitch motion) through modest angles. The patient is asked to look at a fixation light which is located at the position

previously selected in the planning process, and AP and lateral radiographs are taken. The observed clip-to-aperture relationships (panel (m)) can be compared with the computer prescription (panel (l)) and the position and direction of gaze of the patient can be adjusted, if necessary, until agreement is reached. During treatment, the anterior of the patient's eye is viewed on a closed-circuit monitor. Motions of as little as 0.5 mm can be detected and, if necessary, treatment paused until the desired direction of gaze is recovered.

CLINICAL EXPERIENCE WITH PROTONS

I had originally thought to include a short review of the clinical experience to date with protons. I desisted, however, in part because it is such a rapidly changing field that what I write at the time of going to press will be very quickly out of date. But my more important reason is that such a review should be presented by a clinical expert, and should provide all the information that other, one hopes skeptical, clinical experts require to judge the material being presented. I am not qualified to do this.

I do not abandon you entirely in this matter, however. Recently a group of ten papers were published in Acta Oncologica (SPTC, 2005) providing a comprehensive review of the experience in proton beam therapy up to the time of publication. My wife and I were invited to write an editorial (Goitein and Goitein, 2005), and I can best summarize our view of the clinical experience with protons by quoting some of our words:

> These ten papers give a thorough overview of the available clinical data. Taken together with the underlying physical rationale, these data certainly support the proposition that proton beam therapy is a valuable tool in the therapeutic armamentarium. However, so far as clinical results are concerned, while there has been quite a lot of favorable experience, there have been only two randomized studies and very few critical comparisons with historical controls. This is largely due to the fact that, until quite recently, only a few centers have been engaged in proton beam therapy and those that were had limited capacity and a number of constraints such as limited energy, limited technology (e.g., no gantry), and limited beam availability. And, where the initial experience has been very favorable, subsequent randomized trials have not been thought to be possible on ethical grounds. The experience to

date should perhaps be read as a confirmation that the theoretical arguments for proton beam therapy have been upheld in the limited number of situations in which they have been tested.

The physical rationale for proton beam therapy is unimpeachable. Under virtually every scenario, protons deliver less dose outside the target volume than do X-rays – typically they deposit one-half or less integral dose to uninvolved normal tissues than do X-rays (Lomax *et al*., 1999). This statement holds no matter what the technical approach – it is the case, for example, for intensity-modulated proton *vs*. photon therapy. Glimelius *et al*. (2005) […] cite a remarkable 52 published treatment planning comparisons which document this fact. Faced with the possibility of receiving the dose distribution possible through a proton treatment, it is hard indeed to imagine anyone readily volunteering to receive an additional, say, 20 to 30 Gy to a large volume of tissue for which irradiation is not medically indicated.

All this having been said, it is important to appreciate that the application of protons is not without its difficulties and some limitations. With regard to the former, we see it as essential that anyone entering the field of heavy charged particle therapy serve an apprenticeship at one of the existing heavy charged particle centers. The physical/technical limitations include: the management of the influence of internal tissue heterodensities; the substantial problems posed by surgically implanted metallic objects; the lack of superficial skin-sparing; the management of moving target volumes; the unavoidably enlarged penumbra at large depths; the distortion of the dose distribution under conditions of tangential irradiation of structures with strong differences in density (including the skin/air interface); neutron backgrounds which are especially problematic when scattered beams are used in pediatric treatments, and so forth. While many of these limitations can be overcome, nevertheless protons are not uncritically appropriate for all patients. One must always keep in mind that the colorful and attractive pictures produced by treatment planning programs may be misleading.

The commonly raised issue, ultimately, is that of economics. [Most people are persuaded that, if it cost no more than X-ray therapy, protons would in almost all cases be the preferred modality.] Is the drawback of receiving the extra dose delivered by X-rays worth the reduction in cost that they offer? In order to

> answer this question, one has to know or carefully estimate the extent and clinical significance of the benefit, the difference in cost, and how to juxtapose these in a sensible manner. It seems to us that there is doubt on all these matters. Probably the cost issue is the best understood and, in fact, the additional cost of proton beam therapy is not so great as is often thought. Proton beam treatments, by the time a [new] facility is built, will probably cost between 1.7 and 2.1 times the cost of IMRT with X-rays (Goitein and Jermann, 2003). However, it is unclear what the denominator should really be. The cost of some systemic therapies is substantially higher than the cost either of X-rays or protons. When referenced to such costs, the differences in the costs of proton and X-ray therapies are very modest.
>
> Assuming that one knows the cost and the fact that some benefit will accrue from the use of protons, the question remains as to whether the advantage is worthwhile. That is, how high should one set the bar? If one sets it low, then virtually all […] cancer patients requiring radiation therapy would benefit from protons; the higher one sets the bar, the fewer the number of patients one would select to receive protons. Thus, in attempting to make number estimates, this issue is a critical one.

For most of its 40-year history, proton therapy has generally been available only in research institutes, but with its wider availability in hospitals, the range of indications treated will certainly increase. One under-explored area is in the area of the treatment of pediatric cancers. Children's organs are still growing and evolving and this makes them more susceptible to damage by radiation. Due to the reduced doses to all normal tissues, proton therapy will almost certainly make a substantial impact in the treatment of childhood tumors – particularly by reducing untoward side effects. The case is similar for radiation therapy applied concurrently with chemotherapy. Such combined treatments put a significantly greater burden on all normal tissues, potentially making the patient more sensitive to treatment reactions from either or both of the treatment modalities. The reduced doses delivered by protons to the normal tissues could significantly improve the tolerance of patients to such treatments and perhaps allow the use of a greater intensity of one or both of them.

SUMMARY

Protons have come of age as a clinical tool. They have moved from the laboratory to the clinic, and from an obscure activity to a real option for hospitals wishing to provide the most advanced care for their cancer patients. The routine implementation of intensity-modulated proton therapy will place proton beam therapy near the physical limit of its possibilities. Nevertheless, as I have tried to point out, there remain important challenges; these should make proton therapy a fruitful field of research for physicists for some time to come.

The physical characteristics of protons dictate that the total dose delivered to a patient will always be reduced in comparison to photons. Thus, protons will almost always produce a physically superior dose distribution to photons – when a comparable sophistication of technique is used. The very important, and quite controversial, question is whether, and to what extent, this physical improvement will translate into a significant clinical advantage.

12. QUALITY ASSURANCE

This is a very short chapter, but it is nevertheless one of the most important ones. Quality assurance is the process by which one assures oneself that *what is done is that which was intended to be done*. Quality assurance (QA) is the *sine qua non* of any practice, but especially one in which people's lives and health are at stake.

In the matter of quality assurance, the process of radiation therapy is no different from most forms of industrial production – for which there are tried and true methods to achieve and maintain safety and reliability. Safety must be built into the system from the beginning, in the design and construction of all components of the system – as well as in their maintenance and use.

A radiation therapy facility is complex, with very many parts and with complex interactions between them. Faults can arise from a multitude of causes and scenarios, many of which are probably not even known to the user. It is absolutely impossible in practice to test all combinations and permutations of the input signals, so a targeted strategy must be developed to monitor the most important functions and failure modes. This process is termed *quality control* when it is applied to the verification of input signal processing, and *quality assurance* when the performance of the overall system is verified. QA involves a very large number of checks that must be made at various time intervals, typically: in real time (e.g., the ratio of dosimeter responses); daily (e.g., absolute dose delivery, field size, positioning aids); weekly; monthly; and annually.

Whole books are written about QA, but in sum what has to be done is very simple. One must:

1. determine and document, in the case of a procedure, how one plans to go about implementing it – or, in the case of equipment or software, how it should function;
2. decide how to test whether what is done or built is that which was intended to be done or built;
3. perform the tests – and, if unsatisfactory, do whatever is necessary to fix the problem(s) *and then, re-test;*

4. document the tests, corrective actions (if any), and re-tests (if any); and
5. constantly monitor all aspects of steps (1) through (4) to ensure that the quality assurance procedures are actually being followed.

It is essential that a QA program be in place for all elements of the process of planning and delivering radiation therapy. And, it is desirable that the overview of the QA program (step 5) be undertaken by individuals who are not responsible for conducting steps 1 through 4 of the above list.

Some of the more important elements of the radiation therapy process that must be the subject of QA are:

- the integrity and proper functioning of all imaging devices and image data;
- the integrity and proper functioning of all software and associated data - e.g., treatment planning programs;
- the integrity and accuracy of the patient data;
- the integrity and proper functioning of all equipment – e.g., the therapy machine(s);
- the integrity and proper functioning of all equipment-related software – e.g., the control and safety systems, and buried firmware; and
- the proper execution of the established procedures for planning and delivering radiation therapy.

When asked how one assures, to the extent that is possible, the safety and accuracy of radiotherapy, I have one simple answer. "*Quality is assured through meticulous attention to myriads of small details*."

13. CONFIDENCE

INTRODUCTION

Confidence is the flip side of uncertainty. The more uncertain one is about something, the less confident one can be in it. In radiation oncology one is almost always dealing with probabilistic situations. Cure itself is probabilistic; while a given patient will be either 100% or 0% cured, he or she will have *a priori* a probability of cure typical of what was experienced by patients with similar disease treated similarly.

In this Chapter, as well as in Chapter 2, I have avoided encumbering my statements with the many qualifications that they, rigorously speaking, should receive. I have done so in order to focus on concepts, and not to lose sight of them in a cloud of arithmetic and qualifications. I am not implying, by this, that I (or you) can disparage more rigorous treatment of statistical issues. But, unless you are really a crack at statistics, you should try to get the cooperation of a good biostatistician in any research project you

embark on. The pitfalls are many. An expert can help you avoid them, and will also be able to help you get the most out of your data.

LEVELS OF CONFIDENCE

I first want to address the issue of the level of confidence that is necessary in order to make a decision of some sort. I speak here of the confidence one has in a measurement, or in the truth of a hypothesis. It is my thesis that the focus on one particular level of significance can lead to misunderstandings of the value of, and inferences that can or should be made from, data. I think it is important to recognize that data may be used for a number of quite different purposes – and that quite different confidence levels may be appropriate for those purposes. By too heavily emphasizing the 95% confidence level, investigators and readers may unduly restrict the use they make of clinically useful information.

There is, of course, a continuum of levels of confidence. For purposes of discussion, I have selected four and have named them in increasing order of confidence: *hint, trend, statistical significance,* and *conviction*. Although it is out of order, I start with statistical significance.

Statistical significance

Statistical significance is, by convention, considered to be a level of confidence of 95% or better; this corresponds to the P statistic[1] being ≤0.05. This level of confidence has become the benchmark for statistical comparisons in the medical literature.

In the clinical setting, 95% confidence seems to be reasonable as a threshold level for important clinical decisions that affect large numbers of patients. (However, see the discussion below.) Certainly, one standard deviation – about 68% confidence for normally distributed data, or one chance in 3 of being consistent with, say, the null hypothesis – seems too low a level to prompt such action; clinical practice would be too unstable at this level. Three standard deviations – which is equivalent to 99.7% confidence (one chance in 370 that the

[1] The P-value is the probability, assuming the null hypothesis (see below) is true, that the test statistic would take a value as extreme or more extreme than that actually observed.

null hypothesis is correct) – seems too high a level to set the bar; it would make it hard to prompt any changes in practice, given the statistical realities of clinical trials.

Hint

Results that are far from "significant" at the 95% confidence level may nevertheless form the basis for major scientific breakthroughs. Every scientist knows that the faintest clues can stimulate highly productive research. In my opinion, statistically insignificant results can be very important and should not be ignored. The outlier data point, the signal that barely peeks above the background, must always be seriously considered by an investigator; to ignore them is to risk missing clues from nature. Somewhat arbitrarily, I use the term "hint" to describe results between the one and one-and-a-half standard deviation level. Of course, hints occur all the time and, if always pursued, would leave no time for the central research. Progress comes from the combination of noticing such hints and formulating fruitful hypotheses based on the hint *and* a feel for possible underlying mechanisms.

Trend

The term "trend" is commonly used to describe results that, while stronger than hints, do not reach the 95% confidence threshold. The precise meaning of the term is vague; I myself think of a trend as any result that lies between the one-and-a-half and two standard deviation level of significance, which is roughly between a P value of 0.15 and 0.05. When a trend is discovered, one is faced with a vexing clinical problem. Suppose some clinical data indicate that a new therapy cures twice as many patients as the conventional therapy and the statistical analysis rejects the null hypothesis with 85% confidence. This means that there is only a 1 in 7 chance that the new therapy is the same as the conventional therapy – and the most likely difference is the measured difference of a factor of two. Can this be ignored?

Many factors go into deciding on the best therapy for a patient. Certainly the known and, even more so, the unexpected side-effects of treatment and many other issues may qualify the interpretation of the central result. There is a good case for a strong degree of conservatism in clinical practice. It makes sense to require new therapies, as well as diagnostic techniques and other procedures, to stand the test of time, and to be only slowly and carefully instituted. This conservatism lies in part behind the choice of 95% as a threshold

confidence level, as has already been mentioned. Nevertheless, I do not think that doctors can ignore trends in advising their patients about treatments. I address this point in the section below entitled "Randomized Clinical Trials."

A trend becomes more compelling when there is a mechanistic reason to believe it. We often undervalue our understanding of biology. If, for example, there were a widely obeyed dose-effect relationship, and a new technique permitted higher doses to be administered without what was judged to be an appreciable likelihood of additional morbidity, a trend in the results that favored the new technique might well be sufficient to allow its adoption.

One should not forget that trends are also heavy hints.

Conviction

Even a well-designed experiment with a result having statistically clear and unambiguous "significance", say at the $P<0.005$ (three standard deviation) level or more, may be wrong or misleading. Every scientist can tell stories of experiments that gave statistically impeccable results, but that were nevertheless not reproducible. Human error, systematic bias, faulty assumptions, an unrepresentative patient population, multiple comparisons, any one of these can overwhelm the statistics of a clinical trial.[2]

---

[2] I learnt this lesson very early in my professional career. While still a graduate student engaged in an experiment in elementary particle physics, another group in my laboratory made measurements to test the very successful theory of quantum electrodynamics. They observed an apparent violation of the theory and, as soon as their finding became known, some theoreticians proposed that their experiment could be explained if a hitherto unsuspected heavy electron existed. My group was in a position to test this hypothesis immediately, so we were given top priority to use the facility to look for such a high mass electron. To do this we looked at the energy spectrum of particles scattered off a liquid hydrogen target. If there were an excited electron, it would appear as a peak in the energy distribution. And, sure enough, on the very first night, we saw such a peak – standing well above the background. We were of course, thrilled and began to discuss what to name the new particle – and the proper attire to wear in Stockholm. Cooler heads prevailed and we repeated the experiment under different conditions – and the peak disappeared, never to be seen again no matter how hard we tried or under what conditions (including the original one) we measured. Almost certainly our "statistically significant" peak was

[continued on next page

The only secure basis for scientific progress is in the reproduction of results, both by the original investigator and by others. Until several different experiments have given similar answers, it is unwise to be convinced. When they have, conviction may yet be reversed, but it is reasonable. Experiments need to be independently repeated, over and above what would be considered necessary from statistical considerations alone.

Summarizing

In summary, data may be used for a variety of purposes over a wide range of levels of confidence. Most particularly, there should be quite different requirements for entertaining or postulating hypotheses – where hints or trends are quite appropriate and for confirming theories or implementing therapies or tests in general clinical practice where the requirement of 95% confidence or better, and of conviction in the sense used above, is entirely appropriate. We should always bear in mind the level at which a result has been established, and be willing to use it for purposes consistent with the confidence we can properly have in it.

HYPOTHESIS TESTING AND MEASUREMENT

When, say, two therapies are compared, one may interpret the results of an experiment in two rather different ways.

Let us assume that the outcomes in a trial comparing two therapies are $\texttt{result}_1$ and $\texttt{result}_2$. (I leave until later the issue of what is meant by a "result" in this setting.) First, one can consider the experiment as a *measurement* of the difference in outcomes, `diff`, where $\texttt{diff} = \texttt{result}_1 - \texttt{result}_2$. A statistical analysis then estimates the uncertainty in `diff` based on the uncertainties in `result1` (standard deviation $=\texttt{SD}_1$) and `result2` (standard deviation $=\texttt{SD}_2$). This allows one to make a statement of the range of values within which the true value `of diff` is expected to lie, at some stated level of confidence. This is the '*confidence interval*'. At

---

due to some experimental artifact; perhaps to an instrumental failure. Had we not repeated the experiment, we would have been highly embarrassed to have published a statistically significant but quite false result. (Parenthetically, the original observation of a violation of quantum electrodynamics was also wrong and the error was eventually attributed to problems of instrumentation.)

the 95% confidence level (which is about the same as 2 standard deviations) it is roughly the range of values between $2\sqrt{(SD_1^2 + SD_2^2)}$ and $-2\sqrt{(SD_1^2 + SD_2^2)}$. If the value of `diff` lies just outside that range, then there is a 95% probability that `diff` is non-zero – i.e., that the two arms are different. (One of the implicit assumptions in the crude formulae I present here has been that the data being analyzed follow a normal (i.e., Gaussian) distribution.)

However, we have learnt more than that; we have learnt how different they may be. That is, we have *measured* `diff`. This entitles us to say that, there is a 95% likelihood that the true value of `diff` lies between the values $\texttt{diff} - 2\sqrt{(SD_1^2 + SD_2^2)}$ and $\texttt{diff} + 2\sqrt{(SD_1^2 + SD_2^2)}$. Not only may this have excluded the value of zero, it also sets an upper limit on how high `diff` can be, and a lower limit on how low it can be – always at the stated level of confidence. Moreover, in the absence of other information, one's best estimate of the value of `diff` is the value that one measured.

Alternatively, one can consider the experiment as testing the *null hypothesis,* which is the hypothesis that $\texttt{result}_1$ and $\texttt{result}_2$ are measurements of the same quantity. Hypothesis testing asks "if the null hypothesis is true, what is the chance that the measured difference is the result of inevitable statistical fluctuations in measurements of the same quantity?" That is, "what is the level of confidence that, if the true value of `diff` is zero, we would have measured an absolute value of |`diff`| or greater?" The P statistic answers this question. If P=0.05, for example, then there is only a 5% such chance.

If P=0.05, can one then say that, there is only a 5% chance that the null hypothesis is correct? Or, to me equivalently, that there is a 95% chance that the two arms are different? I would say, "rigorously, no, but practically, yes."

Physicists tend to be more comfortable with the concept of measurement; biostatisticians tend to be drawn to tests of the null hypothesis. I think this is because biostatisticians want to ensure that experiments are designed to test a well-defined issue, and to prevent "data dredging." Data dredging is the process of *post hoc* analysis of data, trying to find some pattern in it – for example, to look for some characteristic or combination of characteristics of patients that correlates with some outcome. The problem arises because, even if the data are entirely random, if one looks a sufficient number of times (say, more than 20) for a correlation one is likely eventually to find

one that appears to be established at the 95% confidence level, even though it is really a statistical fluctuation. To avoid this, one should only test the hypothesis that was made before the experiment was run. Personally, while I accept the statistical validity of this attitude, I reject the implication that one should not dredge one's data. Of course one should "listen" to ones data. Of course, one should look for even subtle correlations. The point is that, if one finds something, one has to downgrade its confidence level – even to the point of demoting it to a "hint" that needs to be followed up in a future trial.

There is a very important difference between the hypothesis testing and measurement approaches. Hypothesis testers tend to stop at the point of stating how likely it is that two therapies are equivalent without paying sufficient attention to the magnitude of the difference. With large numbers of subjects one can find a statistically highly significant difference between two therapies that is, nevertheless, too small to be clinically significant. Thus, there is more information in a measurement than in a test of the null hypothesis. For this reason, many biostatisticians these days recommend reporting both statistics.

RANDOMIZED CLINICAL TRIALS: QUANTITATIVE ISSUES

I now want to address the subject of the prospective randomized clinical trial (RCT). Randomized clinical trials have been responsible for important advances in medical care in general, and in radiation oncology in particular. It is by no means my goal to argue against them. However, I want to discuss some of the difficulties in planning and conducting an RCT.

The goal of an RCT is, of course, to determine whether one can say with confidence that one therapy leads to a better outcome than another – and, I might add in view of the discussion above, to measure the quantitative difference between the outcomes, if any. The desire to have randomized trials is based in part on the tendency of randomization to eliminate, or at least substantially reduce, any bias in patient selection. In the context of radiation therapy the prototypical trial would compare a new therapy or a new variant of an existing therapy (the experimental arm) with the best current practice (the control arm). In conducting such a trial it is important to select the outcome of interest. If it were overall survival, for example, then one would be asking one or both of two questions: (a) whether the experimental arm leads to a different survival (e.g., a Kaplan-Meyer

actuarial survival curve) than the control arm; or (b) what the difference in 5-year survival might be.[3]

The combination of local control and morbidity

One often wants to assess the difference in local control for two therapies.[4] But, this seemingly straightforward goal is complicated by the fact that local control is tempered by morbidity. The "goodness" of a therapy is some sort of a combination of the likelihoods of tumor control (TCP) and morbidity – and we often don't really know how to combine these two into a single measure of goodness. Moreover, morbidity is not a singular quantity. The patient is at risk for a variety of complications of different severity, and of variable importance relative to one another and to local tumor control. This is essentially the same problem as has already been discussed in Chapter 9 in the context of the optimization of treatments.

A way out of this conundrum is based on the fact that usually the intensity of a therapy can be adjusted, and the likelihood of morbidity (and of local tumor control) is a function of the treatment intensity. One can then ask for example whether, relative to the control arm, the experimental arm, *adjusted in intensity to give the same likelihood of morbidity as the control arm*, yields a higher probability of local tumor control.

While this is a recognized problem, most RCTs in radiation oncology are nevertheless restricted to two arms, largely to ensure adequate patient accrual. Thus, it is not uncommon to find, after a trial has been concluded, that it is hard to draw useful clinical conclusions because, say, the experimental arm has simultaneously shown an improved TCP and increased morbidity. This is the situation depicted schematically in Figure 13.1a.

---

[3] The P statistic comparing two actuarial survival curves measures the likelihood that their shapes are different, *taken as a whole*. It is important to recognize that two curves might be statistically very different, but have (within stated confidence limits) the same five-year survival. This would perhaps be due to the two therapies having a rather different outcomes in the early years, but leveling out to the same survival level at later times. If one were primarily interested in long term survival, then the P statistic in this case might be somewhat misleading.

[4] For simplicity, I confine my discussions to the comparison of two different therapies. Of course, multiple therapies can be inter-compared in multiple-arm studies.

I want to suggest here that, when patient numbers allow, one should perform RCTs with one control arm and two experimental arms featuring the same therapy but with different intensities. One then has the situation depicted in Figure13.1b. If the intensities of the experimental arms have been judiciously chosen, one may be able to interpolate (or even, modestly extrapolate) between the results of the two experimental arms to estimate the TCP of the experimental arm whose intensity is such that it leads to the same morbidity as the control arm. (Of course, this assumes that a linear interpolation can be made – which, for only modestly different results, is probably a reasonable assumption.) One can then estimate, as depicted in Figure 13.1b, the increase in TCP when the levels of morbidity of the experimental and control arms are the same.

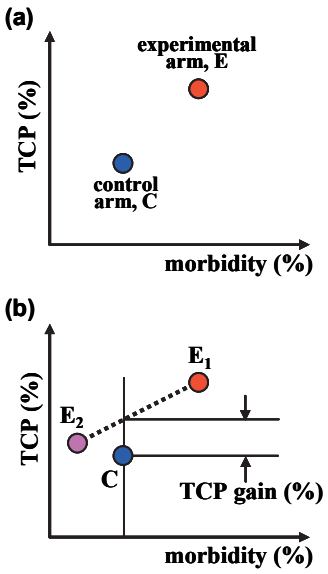


Figure 13.1. Schematic representation of the results of: (a) a standard 2-arm RCT and (b) a 3-arm trial where different intensities of the experimental arms (E1 and E2) are used (see text).

Such an approach must include an estimate of uncertainties, and the process of interpolation or, worse, extrapolation, will magnify these. In consequence, more patients are required for this type of study than for conventional two-arm studies.

RANDOMIZED CLINICAL TRIALS: NON-QUANTITATIVE ISSUES

Finally, I would like to make some brief comments regarding the social aspects of randomized clinical trials. I base these comments on two well-established principles, a commitment to which is, I believe, an important precondition for conducting an RCT and for recommending a patient to enter one.

Equipoise A critical principle in justifying randomization is that one can assure an eligible patient that, given the state of knowledge at the time, it is truly a toss-up as to which arm will be found to be superior, if either. This condition is known as

equipoise and has been well discussed by Hellmann and Hellmann (1991). It should be a required element of an RCT that the arms be in equipoise. If they are not, with few exceptions, the trial would have to be deemed unethical.

It is, of course, rare that investigators do not hope, based on some prior knowledge, that the experimental therapy will be better than the control arm in some important respect, such as the probability of tumor control. However, equipoise can be realized, in this example, if there is a balancing possibility of morbidity the experimental arm. When I refer to the "result" of an arm of a trial, I mean some measure of the overall goodness of the therapy, taking *all* outcomes into account.

The compact with the patient The primary responsibility of a patient's doctor is for the patient's best interest. I believe that patients come to doctors assuming this, and that a doctor's acceptance of a patient is an implicit assurance that this will be the case. They have formed a compact with one another. Of course, doctors have other responsibilities, including those to future patients, but in my view the individual patient's interest must always trump those competing interests.

These days there are numerous pressures to undertake RCTs. Not least of these pressures is the demand of "evidence-based medicine" that new methods have proven their superiority, preferably through the mechanism of RCTs, before they are implemented in routine clinical practice,[5] In some countries and situations, a new therapy cannot be reimbursed in the absence of a supporting RCT. Then, too, there are more theoretical pressures to perform RCTs because they are thought in the academic setting to be rigorous (and publishable). The preceding two principles can, in some circumstances, be in conflict with these pressures.

What the doctor "knows"

Doctors often justify their recommendation that a patient accept randomization on the grounds that they do not "know" which arm is better. They can say this because they are using the word "know" in a

[5] An interesting discussion of the impact of evidence-based medicine on medical practice is to be found in Groupman (2007).

particular and very strong manner, usually in the sense that there is no evidence that the odds are as high or higher than 19:1. The patient, on the other hand, may understand something quite different by the word "know". He or she is very likely to expect the doctor to base his or her recommendation on what he or she "thinks" – that is, to take clinical trends into account and not only statistically significant trials. For example, the patient will probably be well-satisfied if it is the case that comparisons of patients treated using the experimental therapy with historical controls, bolstered perhaps by theoretical considerations, would lead a knowledgeable expert to the conclusion that the odds are in favor of the experimental arm, though by less than a 19:1 margin. For the patient, an odds ratio of 3:1 would quite likely be a sufficient basis for opting for the probably more beneficial arm.

It is widely held that, if a doctor thinks one arm of an RCT may be better than another, he may nevertheless subordinate his belief to that of the "experts" who designed and authorized the trial. He can say with truth that others have decided the arms are in equipoise. For me, however, if his or her own belief is otherwise, the compact with the patient compels the doctor to disclose his or her personal opinion to the patient, who can then make the choice of whether or not to offer himself or herself for randomization in the trial.

The current patient *versus* future patients

The primary reason for conducting an RCT is to gain information that may benefit future patients. This is, I believe, a noble goal. What should one say, however, about a trial in which one arm may be, in one's own opinion, less good than another? May one give what one thinks may be less than optimal therapy to one's own patient in the interests of large numbers of future patients? The answer to this question is influenced by cultural issues. Leaving aside the distressing issue of the conduct of RCTs in third world countries in which informed consent is unusual, there are many developed countries in which the view is taken that the interests of future patients come before those of an individual patient. I have even been told by a European colleague that it would be unethical *not* to treat a patient in the context of an RCT; the greater interests of the population demand it. I myself cannot accept this view. Nor, do I think, could most patients if they were told that this was the attitude with which they were being advised.

There is an important exception to this, however. Patients who are terminally ill are often pleased to be able to contribute to knowledge

from which they are unlikely to benefit and, if this is the case, they should be allowed to do so.

The current patient *versus* current patients

In the section entitled "Statistical significance" near the beginning of this Chapter, I asserted that "95% confidence seems to be reasonable as a threshold level for important clinical decisions that affect large numbers of patients." On the other hand, just above, I asserted that "for the patient, an odds ratio of 3:1 would quite likely be a sufficient basis for opting for the probably more beneficial arm [of a trial]." Can it be the case that the individual patient should be treated differently – on the basis of a hint or trend – while public policy (not to mention reimbursement) should be based on a higher standard – namely that of statistical significance or even conviction?

The arguments presented here would tend to make one think that there is, and should be, a different approach for the individual as compared to the herd. But, I must confess this conclusion leaves me a little uneasy. I think that, if I were a doctor in practice, I might hold the more liberal view. As Minister of Health, I might tend to the conservative approach. (This explains my lack of desire to become a public health expert.)

Scarce resources

Not infrequently the argument is made that the availability of an experimental device or drug is so limited that it cannot be offered to everyone who might need it, and that randomization is then the fairest way of selecting patients to receive the experimental therapy, with the benefit that information useful to others may result. I think this is an acceptable argument *provided that there is no way of selecting those patients who might be expected to derive more benefit than others*. If the resource is truly limited, then randomization could be acceptable under this circumstance.

Continuing an RCT

When should an RCT be stopped? Clearly, a trial will be stopped once the number of patients necessary to test the hypothesis to the desired level of confidence has been entered into the study. A trial must be stopped earlier if a larger than postulated effect is observed in an interim analysis. There is a body of biostatistical literature on the technically difficult subject of how precisely to decide whether to

terminate a study earlier than anticipated and the stopping rules are generally always made explicit and are not infrequently invoked.

But another more vexing problem arises. What should one say to patients who show up late in the study, once a trend has become clear? Can one then assert that the trial is still in equipoise? I don't see how one can.

Two tactics have been employed to get around this difficulty. The first is that the doctor treating the patient is blinded to the results of interim analyses of the study, relying on the study center to monitor the study and terminate it if necessary. In this way, the doctor would not know if the patient were indeed being subjected to poorer, though not "significantly" so, odds. I have a great deal of difficulty with this "solution". It seems to me that consciously avoiding learning information that, if given to the patient, might change his or her mind is incompatible with the trust the patient places in the doctor, and with the patient's legitimate expectation that the doctor be as knowledgeable as reasonably possible, both in general and with respect to the study the doctor is proposing.

The second tactic takes advantage of the fact that, in some studies, the results can only be evaluated long after treatment. All patients necessary to establish statistical significance may then be entered into the study before the results begin to be evident. This seems to me to be more acceptable – although it is still something of a subterfuge, since accrual into the study could be regulated so that interim results could be used to modify and perhaps terminate the study.

Cost-benefit trials

There is a class of RCTs that are aimed, not at the issue of whether one therapy or diagnostic approach is better than another, but on whether the improvement offered by a new and costly therapy or procedure is sufficient to be worth the additional cost. In such trials, there is little doubt that the experimental arm is superior; the trial is designed to measure just how much better it is.

If one were completely open in seeking the patient's informed consent to participate in such a trial, one would have to say to the patient "I want you to take a 50% chance of receiving inferior care so that we can determine whether society can afford to give the better care to everyone in the future." Few patients would agree to do this if it were presented in such a direct manner. This being the case, I think that, generally, such trials should not be attempted.

Summary regarding RCTs

In summary, I believe that the implications of equipoise and of the patient's compact with his or her doctor may in some circumstances preclude randomization. Not every experiment can be done. Ethical considerations may simply make it impossible to conduct RCTs even though they are desirable on purely scientific or public health grounds.

I am concerned that the doctor-patient relationship is at risk in this enterprise, and that the trust the patient places in the doctor to do the best that he or she knows how to do for him or her is at risk of being eroded. There are already many sources of erosion of that confidence, and randomized clinical trials are surely not the most important, but the pressure to perform, promote and participate in randomized clinical trials is, I think, an important source of concern. There is a danger that patients will come to feel that their best interests are not foremost in their doctor's minds, that their doctors are being less than candid with them, or that they have suffered "for science". If this happens it will not be good for doctors, it will not be good for science, and ultimately and most importantly it will not be good for patients.

AFTERWORD

From time to time, the imminent death of radiation oncology is announced, often by advocates of some treatment modality (immunology, gene therapy, and so forth) which is competing for research funds or for "market share." Alas, these obituaries are premature. I say "alas" because we all must hope that some day a more effective approach to the cure of cancer will be discovered. One that will put radiotherapy out of business. A large proportion of my readers will have relatively close family members and friends who have been affected by cancer and they will understand how strong this hope is. Radiation therapy is a blunt and rough tool. It will not turn out to be the ultimate cure. It can, at best, only solve the problem of local, and not metastatic, disease. Its side effects are far from negligible. Our therapeutic gains, the fruit of much hard work over long years, are largely incremental in nature.

I have often been asked by young people contemplating entering the field of Radiation Oncology whether it is not a dead-end field in which employment opportunities and professional satisfaction will dwindle with time. Well, as I said, we hope that this will be so, sometime. But, unfortunately, that time does not seem near. Moreover, even if a highly effective biochemical or other cancer-antagonist is developed, it is likely that, for quite a while, it will be effective vis-à-vis microscopic disease, but not in eradicating the bulk tumor. This is because (1) the sheer burden of tumor cells is likely to be a problem, and (2) the mechanisms for delivery of the agent may be badly compromised in the tumor. For these reasons, it is likely that a tool to sterilize or debulk the gross tumor will be needed for a long time to come, which means that surgery and radiation therapy will continue to play a vital role in the treatment of cancer.

I have often thought that one of the great satisfactions of working in this field is that what one does can make a difference. I think of it as follows. Imagine that there is a universal curve that relates success to intensity of therapy, as in Figure A.1 below. A discipline that lies at a point such as A on the curve, for which one simply could not "get in" enough therapy, would likely be a depressing discipline to practice; the vast majority of one's patients would do poorly. On the other hand, if one's specialty lay at a point such as C, all one's patients

would do well. This would certainly be pleasing, but one might feel that one's patients would have improved without any special effort on one's own part. Radiation oncology more nearly lies at a point such as B. If one is about halfway up the curve, where it is steepest, then one's personal effort has an excellent possibility of improving results. This is, indeed, a charmed situation to be in. Although, one must admit, it has its drawbacks. If one takes credit for successes, then one must be prepared to accept at least partially responsibility for failures.

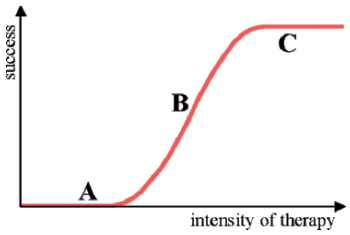


Figure A.1. Possible therapeutic stances (see text).

Several times in a professional lifetime, incautious "experts" are tempted to assert that their given field has reached a point of diminishing returns. That pretty much everything that is to be known ha already been discovered. Don't be deterred by such negativity; there's much to be done. Molecular imaging and targeted therapies may radically change the practice. There is a lot to be gained by a much better understanding of the responses of normal tissues (and tumors) to a whole range of dose–volume distributions. Manipulation of the time factor – the number of fractions, their size, and overall duration of treatment – are important, but poorly understood, variables to be manipulated for the patient's good. Enjoy these opportunities.

Please, resist the ever increasing pressure to be constrained by purely economic considerations. There is no lack of people worrying about finances and figuring out how to cut costs (and corners). Let yourself be an advocate for the patient.

I expect that the role of the individual will become more, rather than less, important. High technology, with good reason, is being brought into the field at an almost alarming rate. However, with increasing complexity and automation come increasing risks. Now, more than ever, both on the physics and medical sides, we need the critical eyes of experts blended with simple common sense to be cast over all that we attempt, and all that we do. The situation glimpsed in Figure 9.10 is not exaggerated; it is a warning – and an opportunity.

All in all, I have found the field of radiation oncology fascinating and personally rewarding, and I myself would have no hesitation to begin

again in these times. This book is written in the hope that it will catalyze or reinforce that same fascination in some of my readers. If you are in the field, or if you plan to enter it, then I'm sure you will have a fruitful, interesting, and enjoyable career.

ACKNOWLEDGEMENTS

We are very fortunate in our profession to be supported by a number of excellent scientific journals. In this connection, I want to mention, *inter alia*: the International Journal of Radiation Oncology, Biology, Physics; Radiotherapy and Oncology; Seminars in Radiation Oncology; Medical Physics; and Physics in Medicine and Biology. I have some idea of the enormous professional effort that the editors and reviewers put into bringing these journals out, month after month, and into keeping them to the highest possible standards. Anyone considering entering the field could do no better than to go to a library and look through a few recent issues of these journals.

A little while ago I went through the pleasant exercise of making a list of the people with whom I had collaborated over the years. I was astounded when I quickly reached the figure of 228. Obviously, it is impossible to recognize so many people individually, and invidious to select only a few. But, I do want to say that I have been enormously fortunate in the colleagues with whom I have worked. They have played a bigger part in writing this book than they know.

I must, however, single out one person, namely Herman Suit. He has been my teacher, colleague and friend for over a third of a century, and I have profited from his knowledge, style, and companionship.

With regard to the preparation of this book, I have had immeasurable help from a number of colleagues. I have bombarded them with questions, begged them for material, and asked them to review part or all of the manuscript. I want, therefore, to express my great gratitude to: Carlo Algranati, Thomas Bortfeld, George Chen, Paul deLuca, Lara Goitein, Bernie Gottschalk, Vincent Gregoire, Eugen Hug, Bleddyn Jones, Norbert Liebsch, Tony Lomax, Alejandro Mazal, Radhe Mohan, Andrzej Niemierko, Harald Paganetti, Eros Pedroni, Marco Schwarz, Steve Selzer, Joel Tepper, Howard Thames, Marcia Urie and Lynn Verhey.

Finally, it is my enormous pleasure to recognize the many contributions of my wife, Gudrun. She has supported me in writing this book at every step, and her review of the manuscript was invaluable to me. It is to her that this book is dedicated.

ACRONYMS

It is, unfortunately, almost impossible to avoid the use of acronyms in a technical field. The following is a list of those used in this book.

"http://physics.nist.gov/cuu/Units/index.html" provides a convenient source of information on SI units.

0D	zero-dimensional (a scalar quantity or number)
1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
3DCRT	three-dimensional conformal radiation therapy
4DCT	3DCT studies repeated at sequential times
A	mass number (no. protons & neutrons in nucleus)
BEV	beam's-eye view
CT	computed tomography
CTV	clinical target volume
DRR	digitally reconstructed radiograph
DVH	dose–volume histogram
EUD	equivalent uniform dose
FSU	functional sub-unit
GTV	gross tumor volume
HU	Hounsfield unit
IM	internal margin
IMPT	intensity-modulated proton therapy
IMRT	intensity-modulated radiation therapy
IMXT	intensity-modulated x-ray therapy
ITV	internal target volume
LET	linear energy transfer ("stopping power")
MLC	multi-leaf collimator

MR	magnetic resonance
MRI	magnetic resonance imaging
MRS	magnetic resonance spectroscopy
NTCP	normal tissue complication probability
OAR	organ at risk
PET	positron emission tomography
POI	point of interest
PRV	planning risk volume
PTV	planning target volume
QA	quality assurance
RBE	relative biological effectiveness
RCT	randomized clinical trial
rf	radio-frequency
RVR	remaining volume at risk
SD	standard deviation (represented by the symbol σ)
SM	setup margin
SOI	surface of interest
TCP	tumor control probability
VOI	volume of interest
WYSIWYG	what you see is what you get
Z	atomic number (no. protons in nucleus)

REFERENCES

Battista J (1980) Computed tomography for radiotherapy planning. Int J Radiat Oncol Biol Phys 6:99–107

Benk VA, Adams JA, Shipley WU, Urie MM, McManus PL, Efird JT, Willett CG and Goitein M (1993) Late rectal bleeding following combined x–ray and proton high dose irradiation for patients with stages T3–T4 prostate carcinoma. Int J Radiat Oncol Biol Phys 26:551–557

Biggs PJ and Shipley WU (1986) A beam width improving device for a 25 MV X ray beam. Int J Radiat Oncol Biol Phys 12:131–5

Bijl HP, van Luijk P, Coppes RP, Schippers JM, Konings AWT and van der Kogel AJ (2003) Unexpected changes of rat cervical spinal cord tolerance caused by inhomogeneous dose distributions. Int J Radiat Oncol Biol Phys 57:274–281

Boon SN, van Luijk P, Böhringer T, Coray A, Lomax A, Pedroni E, Schaffner B and Schippers JM (2000) Performance of a fluorescent screen and CCD camera as a two-dimensional dosimetry system for dynamic treatment techniques. Med Phys 27:2198–2208

Bortfeld T (2003) Physical optimization In: Palta JR and Mackie TR (eds) Intensity Modulated Radiation Therapy: The State of the Art. Medical Physics Publishing, Madison, pp 51–76

Bortfeld T (2006) IMRT: A review and preview. Phys Med Biol 51:R363–R379

Bortfeld T, Biirkelbach J, Boesecke R and Schlegel W (1990) Methods of image reconstruction from projections applied to conformation radiotherapy. Phys Med Biol 35:1423–1434

Bortfeld T, Jokivarsi K, Goitein M, Kung J and Jiang S (2002) Effects of intra–fraction motion on IMRT dose delivery: Statistical analysis and simulation. Phys Med Biol 47:2203–2220.

Bortfeld T, Schmidt–Ullrich R, De Neve W and Wazer DE (2006) Image–Guided IMRT. Springer, Berlin.

Brahme A (1984) Dosimetric precision requirements in radiation therapy. Acta Radiol Oncol 23:379–391

Brahme A (1988) Optimization of stationary and moving beam radiation therapy techniques. Radiother Oncol 12:129–140

Burman C, Kutcher GJ, Emami B and Goitein M (1991) Fitting of normal tissue tolerance data to an analytic function. Int J Radiat Oncol Biol Phys 21:123–135

Chen GT (1988) Dose volume histograms in treatment planning. Int J Radiat Oncol Biol Phys 14:1319–1320

Chen GT, Kung JH and Beaudette KP (2004) Artifacts in computed tomography scanning of moving objects. Semin Radiat Oncol 14:19–26

Chen MF, Lin CT, Chen WC, Yang CT, Chen CC, Liao SK, Liu JM, Lu CS and Lee KD (2006) The sensitivity of human mesenchymal cells to ionizing radiation. Int J Radiat Oncol Biol Phys 66:244–253

Cormack AM (1987) A problem in rotation therapy with X–rays. Int J Radiat Oncol Biol Phys 13:623–630

Deasy JO, Shepard DM and Mackie TR (1997) A proposed delivery method for for conformal proton therapy using intensity–modulation. In: Leavitt DD and Starkshall G (eds) Proc XIIth International Conference on the use of Computers in Radiation Therapy. Medical Physics Publishing, Madison

De Neve W, Wu Y and Ezzell G (2006) Practical IMRT planning. In Bortfeld T, Schmidt–Ullrich R, De Neve W and Wazer DE (eds.) Image–Guided IMRT. Springer, Berlin.

Drzymala R, Mohan R, Brewster L, Chu J, Goitein M, Harms W and Urie M (1991) Dose–volume histograms. Int J Radiat Oncol Biol Phys 21:71–78

Egger E, Zografos L, Schalenbourg A, Beati D, Bohringer T, Chamot L and Goitein G (2003) Eye retention after proton beam radio-therapy for uveal melanoma. Int J Radiat Oncol Biol Phys 55: 867–880

Ellis F (1968) Time, fractionation and dose rate in radiation therapy. In: Vaeth (ed) Frontiers of radiation therapy and oncology. Karger, Basel, vol. 3:pp 131–140

Emami B, Lyman J, Brown A, Coia L, Goitein M, Munzenrider JE, Shank B, Solin LJ and Wesson M (1991) Tolerance of normal tissue to therapeutic irradiation. Int J Radiat Oncol Biol Phys 21:109–122

Glimelius B, Ask A, Bjelkengren G, Björk–Eriksson T, Blomquist E, Johansson B, Karlsson and M, Zackrisson B for the Swedish Proton Therapy Centre Project (2005) Number of patients potentially eligible for proton therapy. Acta Oncologica 44:836–849

Goitein M (1972) Three dimensional density reconstruction from a series of two dimensional projections. Nuclear Instr Methods 101:509–518

Goitein M (1977) The measurement of tissue heterodensity to guide charged particle radiotherapy. Int J Radiat Oncol Biol Phys 3:27–33

Goitein M (1978) A technique for calculating the influence of thin inhomogeneities on charged particle beams. Med Phys 5:258–264.

Goitein M (1983) Non–standard deviations. Med Phys 10:709–711

Goitein M (1985) Calculation of the uncertainty in the dose delivered in radiation therapy. Med Phys 12:608–612.

Goitein M (2005) The cell's–eye view: Assessing dose in four dimensions. Int J Radiat Oncol Biol Phys 62:951–953.

Goitein M and Sisterson JM (1978) The influence of thick inhomogeneities on charged particle beams. Radiat Res 74: 217–230.

Goitein M and Abrams M (1983) Multi–dimensional treatment planning: I. Delineation of anatomy. Int J Radiat Oncol Biol Phys 9:777–787

Goitein M and Miller T (1983) Planning proton therapy of the eye. Med Phys 10:275–283.

Goitein M and Chen GTY (1983) Beam scanning for heavy charged particle radiotherapy. Med Phys 10:831–840

Goitein M and Jermann M (2003) The relative costs of proton and X–ray radiation therapy. Clin Oncol 15:S37–S50

Goitein M and Goitein G (2005) Swedish protons. Acta Oncol 44:793–797

Goitein M, Chen GTY, Ting JY, Schneider RJ and Sisterson JM (1978) Measurements and calculations of the influence of thin inhomogeneities on charged particle beams. Med Phys 5:265–273.

Goitein M, Abrams M, Rowell D, Pollari H and Wiles J (1983) Multi–dimensional treatment planning: II. Beam's eye–view, back projection and projection through CT sections. Int J Radiat Oncol Biol Phys 9:789–797.

Goitein M, Niemierko A and Okunieff P. (1997) The probability of controlling an inhomogeneously irradiated tumor: A stratagem for improving tumor control through partial tumor boosting. In: Proceedings of the 19th LH Gray Conference (Brit J Radiol), pp 25–39.

Goitein M, Lomax AJ and Pedroni ES (2002) Protons in the Treatment of Cancer. Physics Today, Sept issue, pp 45–50

Gottschalk B (2004) In: http://huhepl.harvard.edu/~gottschalk/ File named "pbs.pdf" can be extracted from BGdocs.zip

Gottschalk B, Koehler AM, Schneider RJ, Sisterson JM and Wagner MS (1993) Multiple Coulomb scattering of 160 MeV protons. Nucl Instr Methods Phys Res B74:467–490

Grado GL, Larson TR, Balch CS, Grado MM, Collins JM, Kriegshauser JS, Swanson GP, Navickis RJ and Wilkes MM (1998) Actuarial disease–free survival after prostate cancer brachytherapy using interactive techniques with biplane ultrasound and fluoroscopic guidance. Int J Radiat Oncol Biol Phys 42:289–298

Gragoudas E, Li W, Goitein M, Lane AM, Munzenrider JE and Egan KM (2002) Evidence–based estimates of outcome in patients irradiated for intraocular melanoma. Arch Ophthalmol 120:1665–1671

Groupman J (2007) How doctors think. Houghton Mifflin, Boston

Hall EJ (2000) Radiobiology for the radiologist, fifth edn. Lippincot, Williams and Wilkins, Philadelphia

Hall EJ (2003) The bystander effect. Health Phys 85:31–35

Hall EJ (2006) Intensity–modulated radiation therapy, protons, and the risk of second cancers. Int J Radiat Oncol Biol Phys 65:1–7.

Hellmann S and Hellmann DS, (1991) Of mice but not men: problems of the randomized clinical trial. New Engl J Med 324:1585–1589

Herring DF and Compton DMJ (1971) The degree of precision required in the radiation dose delivered in cancer radiotherapy. In: Glicksman AS *et al* (eds.) Br J Radiol special report number 5: Computers in radiation therapy. London, Brit Inst Radiol, pp 51–58

Hong L, Goitein M, Bucciolini M and Comiskey R (1996) A pencil beam algorithm for proton dose calculation. Phys Med Biol 41:1305–1330

Huff CA, Matsui WH, Smith BD and Jones RJ /2006) Strategies to eliminate cancer stem cells: clinical implications. Europ J Cancer 42:1293–1297

Hunt MA, Hsiung CY, Spüirou SV, Chui CS, Amols HI and Ling CC (2002) Evaluation of concave dose distributions created using an inverse planning system. Int J Radiat Oncol Biol Phys 54:953–962

IAEA (2000) International Atomic Energy Agency. Absorbed dose determination in external beam radiotherapy: An international code of practice for dosimetry based on standards of absorbed dose to water, Technical Reports Series No. 398 (International Atomic Energy Agency, Vienna). Revised and updated version (V.11b, 23 April, 2004) available on website: http://www–naweb.iaea.org/nahu/dmrp/pdf_files/COPV11b.pdf.

ICRU50 (1993). International Commission on Radiation Units and Measurements, Prescribing, Recording, and Reporting Photon Beam Therapy, ICRU Report 50. International Commission on Radiation Units and Measurements, Washington

ICRU62 (1999). International Commission on Radiation Units and Measurements, Prescribing, Recording, and Reporting Photon Beam Therapy, Supplement to ICRU Report No. 50, ICRU Report 62. International Commission on Radiation Units and Measurements, Washington

ICRU71 (2005) International Commission on Radiation Units and Measurements, Prescribing, Recording, and Reporting Electron Beam Therapy, ICRU Report 71. International Commission on Radiation Units and Measurements, Washington

ICRU78 (2007) International Commission on Radiation Units and Measurements, Prescribing, Recording, and Reporting Proton Beam Therapy, ICRU Report 78. International Commission on Radiation Units and Measurements, Washington

ISO (1995) Guide to the Expression of Uncertainty in Measurement. International Organization for Standardization, Geneva

Jackson A (2001) Partial irradiation of the rectum. Semin Radiat Oncol 11:215–223

Jaffrey DA (2003) X–ray–guided IMRT. In: Palta JR and Mackie TR (eds) Intensity modulated radiation therapy: The state of the art. Medical Physics Publishing, Madison

Johns HE and Cunningham JR (1983) The physics of radiology, fourth edn. Charles C. Thomas, Springfield

Kartha PK, Chung–Bin A, Wachtor T and Hendrickson FR (1975) Accuracy in patient setup and its consequence in dosimetry. Med Phys 2:331–2

Karzmark CJ and Rust DC (1972) Radiotherapy treatment simulators and automation. A case for their provision from a cost viewpoint. Radiology 105:157–161

Kessler M (2006) Image registration and data fusion in radiation therapy. Br J Radiol 79(1):S99–S108

Khan FM (2003) The physics of radiation therapy, third edn. Lippincott, Williams and Wilkins, Philadelphia

Kjellberg RN, AM Koehler, *et al*. (1962) Stereotaxic instrument for use with the Bragg peak of a proton beam. Confin Neurol 22:183–189

Klein EE, Drzymala RE, Purdy JA and Michalski J (2005) Errors in radiation oncology: a study in pathways and dosimetric impact. J. Appl Clin Med Phys 6:81–94.

Koehler AM (1968). Proton radiography. Science 160:303.

Langen KM and Jones DTL (2001) Organ motion and its management. Int J Radiat Oncol Biol Phys 50:265–278

Liao ZX and Travis EL (1994) Unilateral nephrectomy 24 hours after bilateral kidney irradiation reduces damage to function and structure of remaining kidney. Radiat Res 139:290–299

Liao ZX, Travis EL and Tucker SL (1995) Damage and morbidity from pneumonitis after irradiation of partial volumes of mouse lung. Int J Radiat Oncol Biol Phys 32:1359–1370

Ling CC, Humm J, Larson S, Amols H, Fuks Z, Leibel S, Koutcher JA. (2000) Towards multidimensional radiotherapy (MD–CRT): biological imaging and biological conformality. Int J Radiat Oncol Biol Phys 47:551–60

Ling CC, Yorke E, Amols H, Mechalakos J, Erdi Y, Leibel S, Rosenzweig K and Jackson A (2004) High–tech will improve radiotherapy of NSCLC: a hypothesis waiting to be validated. Int J of Radiat Oncol Biol Phys 60:3–7

Lomax AJ, Bortfeld T, Goitein G, Debus J, Dykstra C, Tercier PA, Couke PA and Mirimanoff RO (1999) A treatment planning inter–comparison of protons and intensity–modulated photon therapy. Radiother Oncol 51:257–271.

Mackie TR, Holmes T, Swerdloff S, Reckwerdt P, Deasy JO, Yang J, Paliwal B and Kinsella T (1993) Tomotherapy: A new concept for the delivery of conformal therapy using dynamic collimation. Med Phys 20:1709–1719

Maintz JB and Viergever MA (1998) A survey of medical image registration. Med Image Anal 2:1–36

Mould RF (1988) Introductory medical statistics. Institute of Physics Publishing, Bristol

Niemierko A (1992) Random search algorithm (RONSC) for optimization of radiation therapy with both physical and biological end points and constraints. Int J Radiat Oncol Biol Phys 23:89–98.

Niemierko A (1997) Reporting and analyzing dose distributions: a concept of equivalent uniform dose. Med Phys 24:103–110

Niemierko A (1999) A generalized concept of equivalent uniform dose (EUD). Med Phys 26:110

Niemierko A and Goitein M (1991) Calculation of normal tissue complication probability and dose–volume histogram reduction schemes for tissues with a critical element architecture. Radiother Oncol 20:166–76.

Niemierko A and Goitein M (1993a) Modeling of normal tissue response to radiation: the critical volume model. Int J Radiat Oncol Biol Phys 25:135–145

Niemierko A and Goitein M (1993b) Implementation of a model for estimating tumor control probability for an inhomogeneously irradiated tumor. Radiother Oncol 29:140–147

Niemierko A and Goitein M (1994) Dose–volume distribution (DVD's): A new approach to dose–volume histograms in three–dimensional treatment planning. Med Phys 21:3–11

OED (2001) New Oxford dictionary of English. Oxford University Press, Oxford

Ohara K, Okumura T, Akisada M, Inada T, Mori T, Yokota H and Calaguas MJ (1989) Irradiation synchronized with respiration gating Int J Radiat Oncol Biol Phys 17:853–857

Paganetti H (2006) Monte Carlo calculations for absolute dosimetry to determine output factors for proton therapy fields. Phys Med Biol:51, 2801–2812

Paganetti H, Niemierko A, Ancukiewicz M, Gerweck LE, Loeffler JS, Goitein M and Suit HD (2002) RBE values for proton beam therapy. Int J Radiat Oncol Biol Phys 53:407–421

Paganetti H, Jiang HV and Trofimov A (2005) 4D Monte Carlo simulation of proton beam scanning: modeling of variations in time and space to study the interplay between scanning pattern and time–dependent patient geometry. Phys Med Biol 50:983–990

Palter JR and Mackie TR (eds.) Intensity–modulated radiation therapy: the state of the art. Medical Physics Publishing, Madison.

Pedroni E (1981) The planning system for the SIN pion therapy facility. In: Burger G and Broerse JJ (eds) Treatment planning for external beam therapy with neutrons. Urban and Schwartzenburg, Munich, pp 60–69

Pedroni E, Scheib S, Böringer T, Coray A, Grossman M, Lin S and Lomax A (2005) Experimental characterization and physical modeling of the dose distribution of scanned proton pencil beams. Phys Med Biol 50:541–561

Pelizzari CA, Chen GT, Spelbring DR and Weichselbaum RR (1989) J Comput Assist Tomog 13:20–26

Powers WE, Kinzie JJ, Demidecki AJ, Bradfield JS and Feldman A (1973) A new system of field shaping for external–beam radiation therapy. Radiology 108:407–411.

Press WH, Flannery BP, Teukolsky SA and Vetterling WT (1988) Numerical recipes in C. Cambridge university Press, Cambridge

Rabinowitz I, Broomberg J, Goitein M, McCarthy K and Leong J (1985) Accuracy of radiation field alignment in clinical practice. Int J Radiat Oncol Biol Phys 11:1857–1867.

Rutz HP and Lomax AJ (2005) Donut-shaped high dose configuration for proton beam radiation therapy, Strahlenther Onkol 181:49–53

Schaffner B and Pedroni E (1998) The precision of proton range calculations in proton radiotherapy treatment planning: Experimental verification of the relation between CT–HU and proton stopping power. Phys Med Biol 43:1579–1592.

Schneider U and Pedroni E (1995) Proton Radiography as a tool for quality control in proton therapy. Med Phys 22:353–363

Schneider U, Agosteo S, Pedroni E *et al*. (2002) Secondary neutron dose during proton therapy using spot scanning. Int J Radiat Oncol Biol Phys 53:244–251

Schweikard A, Shiomi H and Adler J (2004) Respiration tracking in radiosurgery. Med Phys 31:2738–2741

Seltzer S (1993) National Institute of Standards and Technology (NIST) technical note NISTIR 5221

Seminars in Radiation Oncology (2001) Partial Organ Irradiation. Ten Haken RK (ed.) 11(3).

Shalev S, Bartel L, Therrien P, Hahn P and Carey M (1988) The objective evaluation of alternative treatment plans: I. Images of regret. Int J Radiat Oncol Biol Phys 15:763–767.

Shipley WU, Tepper JE, Prout GR, Verhey LJ, Mendiondo OA, Goitein M, Koehler AM and Suit HD (1979) Proton radiation as boost therapy for localized prostatic carcinoma. JAMA 241:1912–1915

Soares HP, Kumar A, Daniels S, Swann S, Cantor A, Hozo I, Clark M, Serdarevic F, Gwede C, Trotti A and Djulbegovic B (2005) Evaluation of new treatments in radiation oncology: Are they better than standard treatments? JAMA 293:970–978

SPTC (2005) Papers from the Swedish proton therapy center investigation. Acta Oncologica 44:836–920

Tepper J (1981) Clonogenic potential of human tumors. A hypothesis. Acta Radiol Oncol. 20:283–8.

Terahara A, Niemierko A, Goitein M, Finkelstein D, Hug E, Liebsch N, O'Farrel D, Lyons S and Munzenrider J (1999) Analysis of the relationship between tumor dose inhomogeneity and local control in patients with skull base chordomas. Int J Radiat Oncol Biol Phys 45:351–358

Tourovsky A, Lomax A.J, Schneider U and Pedroni E (2005) Monte Carlo dose calculations for spot scanned proton therapy. Phys Med Biol 50:971–981

Tufte ER (1990) Envisioning information. Graphics Press, Cheshire, Connecticut

Tufte ER (1997) Visual explanations. Graphics Press, Cheshire, Connecticut

Tufte ER (2001) The visual display of quantitative information, second edn. Graphics Press, Cheshire, Connecticut

Urie M, Goitein M and Wagner M (1984) Compensating for heterogeneities in proton radiation therapy. Phys Med Biol 29:553–566

Urie M, Goitein M, Holley WR and Chen GTY (1986a) Degradation of the Bragg peak due to inhomogeneities. Phys Med Biol 31:1–15

Urie MM, Sisterson JM, Koehler AM, Goitein M and Zoesman J (1986b) Proton beam penumbra: effects of separation between patient and beam modifying devices. Med Phys 13:734–741

Urie M, Goitein M, Doppke K, Kutcher G, LoSasso T, Mohan R, Munzenrider JE, Sontag M and Wong J (1991) The role of uncertainty analysis in treatment planning. Int J Radiat Oncol Biol Phys 21:91–107

van Herk M, Remeijer P, Rasch C and Lebesque JV (2000) The probability of correct target dosage: Dose–population histograms for deriving treatment margins in radiotherapy. Int J Radiat Oncol Biol Phys 47:1121–1135

van Herk M (2004) Errors and margins in radiotherapy Semin Radiat Oncol 14:52–64

van Luijk P, Novakova–Jiresova A, Faber H, Schippers JM, Kampinga HH, Meertens H and Coppes RP (2005) Radiation damage to the heart enhances early radiation–induced lung function loss. Cancer Res 65:6509–6511

Verhey LJ, Koehler AM, McDonald JC, Goitein M, Ma IC, Schneider RJ and Wagner M (1979) The determination of absorbed dose in a proton beam for purposes of charged particle radiation therapy. Radiat Res 79:34–54.

Verhey LJ, Goitein M, McNulty P, Munzenrider JE and Suit HD (1982) Precise positioning of patients for radiation therapy. Int J Radiat Oncol Biol Phys 8:289–294

Verhey L and Bentel G (1999) Patient immobilization. In: Van Dyk J (ed) A Modern Technology of Radiation Oncology. Medical Physics Publishing, Madison, pp 53–94

Viola P and Wells WM III (1995) Alignment by maximization of mutual information. In: Grimson E, Shafer S, Blake A, Sugihara K (eds.) International Conference on Computer Vision, IEEE Computer Society Press, Los Alamitos, CA, pp 16–23

Wang H, Dong L, Lii MF, Lee AL, de Crevoisier R, Mohan R, Cox JD, Kuban DA and Cheung R (2005) Implementation and validation of a three–dimensional deformable registration algorithm for targeted prostate cancer radiotherapy. Int J Radiat Oncol Biol Phys 61:725–735

Waschek T, Levegrün S, Van Kampen M, *et al*. (1997) Determination of target volumes for three–dimensional radiotherapy of cancer patients with a fuzzy system. Fuzzy Sets and Systems, 89:361–370

Webb S and Nahum AE (1993) A model for calculating tumor control probability in radiotherapy including the effects of inhomogeneous distributions of dose and clonogenic cell density. Phys Med Biol 38:653–666

Webb S and Lomax A (2001) There is no IMRT? Phys Med Biol 46:L7–L8

Wilson RR (1946) Radiological use of fast protons. Radiology 47:487–491

Wilson R and Crouch EAC (2001) Risk–benefit analysis. Harvard University Press, Cambridge

Withers HR, Taylor JMG and Maciejewski B (1998) Treatment volume and tissue tolerance. Int J Radiat Oncol Biol Phys 14:751–759

York ED (2003) Biological indices for evaluation and optimization of IMRT. In: Palta JR and Mackie TR (eds) Intensity–Modulated Radiation Therapy. Medical Physics Publishing, Madison

INDEX

D

E

F

L

M

Q

R

S

T