Nuclear Medicine Technology and Techniques

Johan Nuyts

Nuclear Medicine, K.U.Leuven
U.Z. Gasthuisberg, Herestraat 49
B3000 Leuven
tel: 016/34.37.15
e-mail: Johan.Nuyts@uz.kuleuven.ac.be

Februari, 2000

Contents

1 Introduction **2**

2 Radionuclides **3**
2.1 Radioactive decay modes . . . 3
2.2 Statistics . . . 5

3 Interaction of photons with matter **7**
3.1 Photo-electric effect . . . 7
3.2 Compton scatter . . . 8
3.3 Attenuation . . . 8

4 Data acquisition **10**
4.1 Detecting the photon . . . 10
4.1.1 Scintillation crystal . . . 10
4.1.2 Photo Multiplier Tube . . . 11
4.1.3 Two-dimensional gamma detector . . . 12
4.1.4 Resolution . . . 14
4.1.5 Storing the data . . . 15
4.1.6 Alternative designs . . . 15
4.2 Collimation . . . 16
4.2.1 Mechanical collimation in the gamma camera . . . 18
4.2.2 Electronic and mechanical collimation in the PET camera . . . 20
4.3 Energy windowing . . . 27
4.4 Corrections . . . 29
4.4.1 Linearity correction . . . 30
4.4.2 Energy correction . . . 31
4.4.3 Uniformity correction . . . 31
4.4.4 Dead time correction . . . 32
4.4.5 Random coincidence correction . . . 35

5 Image formation **36**
5.1 Introduction . . . 36
5.2 Planar imaging . . . 37
5.3 2D Tomography . . . 39
5.3.1 2D filtered backprojection . . . 39
5.3.2 Iterative Reconstruction . . . 44
5.3.3 Regularization . . . 49

5.3.4 Convergence . . . 50
5.4 Fully 3D Tomography . . . 51

6 The transmission scan **53**
6.1 System design . . . 53
6.2 Attenuation correction . . . 54

7 Quality control **57**
7.1 Gamma camera . . . 57
7.1.1 Planar imaging . . . 58
7.1.2 Whole body imaging . . . 63
7.1.3 SPECT . . . 64
7.2 Positron emission tomograph . . . 66
7.2.1 Evolution of blank scan . . . 66
7.2.2 Calibration . . . 66
7.2.3 Normalization . . . 67

8 Image analysis **69**
8.1 Standardized Uptake Value . . . 69
8.2 Tracer kinetic modeling . . . 70
8.2.1 Introduction . . . 70
8.2.2 The compartmental model . . . 71
8.3 Image quality . . . 77
8.3.1 Subjective evaluation . . . 77
8.3.2 Task dependent evaluation . . . 78
8.3.3 Continuous and digital . . . 78
8.3.4 Bias and variance . . . 78
8.3.5 Evaluating a new algorithm . . . 79

9 Biological effects **81**

10 Appendix **83**
10.1 Poisson noise . . . 83
10.2 Convolution . . . 84
10.3 Combining resolution effects: convolution of two Gaussians . . . 85
10.4 Error propagation . . . 86
10.4.1 Sum or difference of two independent variables . . . 86
10.4.2 Product of two independent variables . . . 86
10.4.3 Any function of independent variables . . . 87
10.5 Expectation of Poisson data contributing to a measurement . . . 87
10.6 The convergence of the EM algorithm . . . 89
10.7 The Laplace transform . . . 90

Chapter 1

Introduction

The use of radioactive isotopes for medical purposes has been investigated since 1920, and since 1940 attempts have been undertaken at imaging radionuclide concentration in the human body.

In the early 1950s, Ben Cassen introduced the rectilinear scanner, a "zero-dimensional" scanner, which (very) slowly scanned in two dimensions to produce an two-dimensional image of the radionuclide concentration in the body. In the late 1950s, Hal Anger developed the first "true" gamma camera, introducing an approach that is still being used in the design of virtually all modern camera's: the Anger scintillation camera [1], a 2D planar detector to produce a 2D projection image without scanning.

The Anger camera can also be used for tomography. The projections images can then be used to compute the original spatial distribution of the radionuclide within a slice or a volume. Already in 1917, Radon published the mathematical method for reconstruction from projections, but only in the 1970s, the method was introduced in medical applications, first in CT and next in nuclear medicine imaging. At the same time, iterative reconstruction methods were being investigated, but the application of those methods had to wait for sufficient computer power till the 1980s.

The Anger camera is often called gamma camera, because it detects gamma rays. When it is designed for tomography, it is also called a SPECT camera. SPECT stands for Single Photon Emission Computed Tomography and contrasts with PET, i.e. Positron Emission Tomography, which detects photon pairs. Anger showed that two scintillation camera's could be combined to detect photon pairs originating after positron emission. Ter-Pogossian et al. built the first dedicated PET-system in the 1970s, which was used for phantom studies. Soon afterwards, Phelps, Hoffman et al built the first PET-scanner (also called PET-camera) for human studies [2]. Since its development, the PET-camera has been regarded nearly exclusively as a research system. Only since about 1995, it is really breaking through as a clinical instrument.

Below, PET and SPECT will be discussed together since they have a lot in common. However, there are also important differences. One is the cost price: PET systems are about 4 times as expensive as gamma cameras. In addition, many PET-tracers have a very short half life, so it is mandatory to have a small cyclotron, a laboratory and a radiopharmacy expert in the close neighborhood of the PET-center.

Chapter 2

Radionuclides

In nuclear medicine, a tracer molecule is administered to the patient, usually by intravenous injection. A tracer is a particular molecule, carrying an unstable isotope, a radionuclide. The molecule is recognized by the body, and will get involved in some metabolic process. The unstable isotopes produce gamma rays, which allow us to measure the concentration of the tracer molecule in the body as a function of position and time. A tracer is always administered in very low amounts, such that the natural process being studied is not affected by the tracer.

Consequently, nuclear medicine is all about *measuring function or metabolism*, in contrast to many other modalities including CT, MRI and echography, which mainly perform anatomical measurements. This boundary is not strict, though: CT, MRI and ultrasound imaging allow functional measurements, while nuclear medicine instruments also provide some anatomical information. However, nuclear medicine techniques provide concentration measurements with a sensitivity that is orders of magnitude higher than that of any other modality.

2.1 Radioactive decay modes

The radionuclides emit electromagnetic rays during radioactive decay. These rays are called "gamma rays" or "X-rays". Usually, electromagnetic rays originating from nuclei are called gamma rays, while those from atoms are called x-rays. However, when they have the same frequency they are indistinguishable, so the distinction is not important.

There are many ways in which an unstable isotope can decay. Depending on the decay mode, one or two gamma rays will be emitted in every event.

1. β^- emission.

   - In this process, a neutron is transformed into a proton and an electron (called a β^- particle). Also a neutrino (ν) is produced and (since the decay result is more stable) some energy is released:

     $$n \rightarrow p^+ + e^- + \nu + \text{kinetic energy}. \tag{2.1}$$

     Since the number of protons is increased, this transmutation process corresponds to a rightward step in Mendeleev's table.

- The resulting daughter product of the above transmutation can also be in an excited state, in which case it rapidly decays to a more stable nuclear arrangement, releasing the excess energy as one or more γ photons. The nucleons are unchanged, so there is no additional transmutation in decay from excited to ground state.
- The daughter nucleus of the decay process may also be in a metastable or isomeric state. In contrast to an excited state, the life time of a metastable state is much longer. When it finally decays, it does so by emitting a photon. In many metastable isotopes, this photon is immediately absorbed by an electron of the very same atom, as a result of which that electron is ejected. This is called a conversion electron.
  The most important single photon isotope ^{99m}Tc is an example of this mode. ^{99m}Tc is a metastable daughter product of ^{99}Mo (half life 66 hours). ^{99m}Tc decays to ^{99}Tc (half life 6 hours) by emitting a photon of 140 keV.

2. Electron capture (EC).

   - An orbital electron is captured, and combined with a proton to produce a neutron:

     $$p^+ + e^- \rightarrow n + \nu \tag{2.2}$$

     As a result of this, there is an orbital electron vacancy. An electron from a higher orbit will fill this vacancy, emitting the excess energy as a photon. Note that EC causes transmutation towards the leftmost neighbor in Mendeleev's table.
   - If the daughter product is in an excited state, it will further decay towards a state with lower energy by emitting additional energy as γ photons (or conversion electrons).
   - Similarly, the daughter product may be metastable, decaying via isomeric transition after a longer time.

3. Positron emission (β^+ decay).

   - A proton is transformed into a neutron and a positron (or anti-electron):

     $$p^+ \rightarrow n + e^+ + \nu + \text{kinetic energy}. \tag{2.3}$$

     After a very short time the positron will hit an electron and annihilate (fig. 2.1). The mass of the two particles is converted into energy, which is emitted as two photons. These photons are emitted in opposite directions. Each photon has an energy of 511 keV (the rest mass of an electron or positron).
   - Again, the daughter nucleus may also be in an excited or a metastable state.

As a rule of thumb, light atoms tend to emit positrons, heavy ones tend to prefer other modes, but there are exceptions. The most important isotopes used for positron emission imaging are ^{11}C, ^{13}N, ^{15}O and ^{18}F. Since these atoms are very frequent in biological molecules, it is possible to make a radioactive tracer which is chemically identical to the target molecule, by substitution of a stable atom by the corresponding radioactive isotope. These isotopes are relatively short lived: ^{11}C: 20 min, ^{13}N: 10 min, ^{15}O: 2 min and ^{18}F: 2 hours. As a

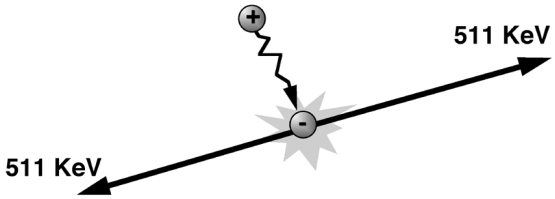

Figure 2.1: *Positron-electron annihilation.*

result, except for ^{18}F, they must be produced close to the PET-system immediately before injection. To produce the isotope, a small cyclotron is required. The isotope must be rapidly incorporated into the tracer molecule in a dedicated radiopharmacy laboratory.

In nuclear medicine, we use radioactive isotopes that emit photons with an energy between 80 and 300 keV for single photon emission, and 511 keV for positron emission. The energy of the emitted photons is fixed: each isotope emits photons with one or a few very sharp energy peaks. If more than one peak is present, each one has its own probability which is a constant for that isotope. So if we administer a tracer to a patient, we know exactly what photons we will have to measure.

2.2 Statistics

The exact moment at which an atom will decay cannot be predicted. All that is known is the probability that it will decay in the next time interval dt. This probability is αdt, where α is a constant for each isotope. So if we have N radioactive photons at time t_0, we expect to see a decrease dN in the next interval dt of

$$dN = -N\alpha dt. \tag{2.4}$$

Integration over time yields

$$N(t) = N(t_0)e^{-\alpha(t-t_0)}. \tag{2.5}$$

This is what we *expect*. If we actually measure it, we may obtain a different value, since the process is statistical. As shown below, the estimate will be better for larger N.

The half life of a tracer is the amount of time after which only half the amount of radioactivity is left. It is easy to compute it from equation (2.5):

$$t_{\frac{1}{2}} = \frac{\ln 2}{\alpha}. \tag{2.6}$$

The source strength used to be expressed in Curie (Ci), but the preferred unit is now Becquerel (Bq). One Bq means 1 expected event per s. For the coming years, it is useful to know that

$$1 \text{ mCi} = 37 \text{ MBq}. \tag{2.7}$$

Typical doses are in the order of 10^2 Mbq. (Marie and Pierre Curie and Antoine Becquerel received the Nobel prize in 1903, for their discovery of radioactivity in 1896).

As shown in appendix 10.1, the probability of measuring r photons, when r photons are expected equals

$$p_r(n) = \frac{e^{-r}r^n}{n!} \tag{2.8}$$

$$= e^{-r}\frac{r}{1}\frac{r}{2}\frac{r}{3}\ldots\frac{r}{n} \tag{2.9}$$

This is a Poisson distribution with r the average number of expected photons. For large r, it can be well approximated by a Gaussian with mean r and standard deviation $\sqrt{r}$:

$$p_r(n) \simeq \frac{1}{\sqrt{2\pi r}}\exp\left(\frac{-(n-r)^2}{2r}\right) \tag{2.10}$$

For smaller ones (less than 10 or so) it becomes markedly asymmetrical, since the probability is always 0 for negative values.

Note that the distribution is only defined for integer values of n. This is obvious, because one cannot detect partial photons (r is a real number, because the average number of expected photons does not have to be integer). Summing over all n values yields

$$\sum_0^\infty p_r(n) = e^{-r}\sum_0^\infty \frac{r^n}{n!} = e^{-r}e^r = 1. \tag{2.11}$$

As with a Gaussian, r is not only the mean of the distribution, it is also the value with highest probability. The signal-to-noise ratio (SNR) then becomes

$$SNR = \frac{r}{\sqrt{r}} = \sqrt{r}. \tag{2.12}$$

Hence, if we measure the amount of radioactivity, the SNR becomes larger if we measure longer.

The only assumption made in the derivation in (appendix 10.1) was that the probability of an event was constant in time. It follows that "thinning" a Poisson process results in a new Poisson process. With thinning, we mean that we randomly accept or reject events, using a fixed acceptance probability. If we expect N photons, and we randomly accept a fraction f, then the expected number of accepted photons is fN. Since the probability of surviving the whole procedure (original process followed by selection) has a fixed probability, the resulting process is still Poisson.

Chapter 3

Interaction of photons with matter

In nuclear medicine, photon energy ranges roughly from 60 to 600 keV. For example, ^{99m}Tc has an energy of 140 keV. This corresponds to a wave length of 9 pm and a frequency of 3.4 $\times$ 10^{19} Hz. At such energies, γ photons behave like particles rather than like waves.

At these energies, the dominating interactions of photons with matter are photon-electron interactions: Compton scatter and photo-electric effect. As shown in figure 3.1, the dominating interaction is a function of energy and electron number. For typical nuclear medicine energies, Compton scatter dominates in light materials (such as water and human tissue), and foto-electric effect dominates in high density materials. Pair production (conversion of a photon in an electron and a positron) is excluded for energies below 1022 keV, since each of the particals has a mass of 511 keV. Rayleigh scatter (absorption and re-emission of (all or a fraction of) the absorbed energy as a photon in a different direction) has a low probability.

3.1 Photo-electric effect

A photon "hits" an electron. The electron absorbs all the energy of the photon. If that energy is higher than the binding energy of the electron, the electron escapes from its atom. Its kinetic energy is the difference between the absorbed energy and the binding energy. In a

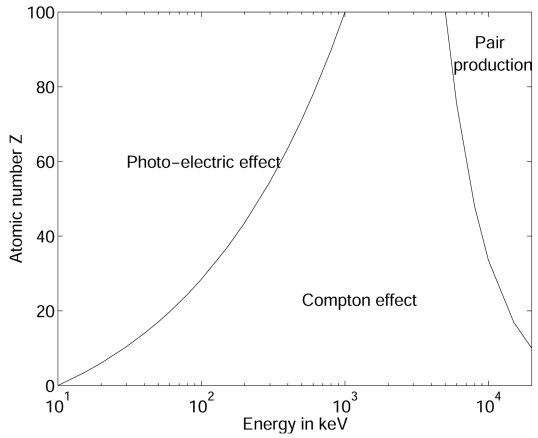

Figure 3.1: *Dominating interaction as a function of electron number Z and photon energy.*

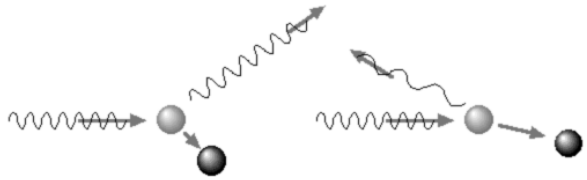

Figure 3.2: *Compton scatter can be regarded as an elastic collision between a photon and an electron.*

dense material, this photo-electron will collide with electrons from other atoms, and will lose energy in each collision, until no kinetic energy is left.

As a result, there is now a electron vacancy in the atom, which may be filled by an electron from a higher energy state. The difference in binding energy of the two states must be released. This can be done by emitting a photon. Alternatively, the energy can be used by another electron with low binding energy to escape from the atom.

In both cases, the photon is completely eliminated.

3.2 Compton scatter

Compton scatter can be regarded as an elastic collision between a photon and an electron. The term "elastic" means that total kinetic energy before and after collision is the same. As in any collision, the momentum is preserved as well. Applying equations for the conservation of momentum and energy (with relativistic corrections) results in the following relation between the photon energy before and after the collision:

$$E'_\gamma = E_\gamma \frac{1}{1 + E_\gamma(1 - \cos\theta)/(m_e c^2)} \tag{3.1}$$

with E'_γ = energy after collision
E_γ = energy before collision
m_e = electron rest mass
θ = scatter angle

The expression $m_e c^2$ is the energy available in the mass of the electron, which equals 511 keV. For a scatter angle $\theta = 0$, the photon loses no energy and it is not deviated from its original direction: nothing happened. The loss of energy increases with θ and is maximum for an angle of 180°. The probability that a photon will interact at all with an electron depends on the energy of the photon. If the interaction takes place, each scatter angle has its own probability.

3.3 Attenuation

The *linear attenuation coefficient* μ is defined as the probability of interaction per unit length (unit: cm^{-1}). Figure 3.3 shows the mass attenuation coefficients as a function of energy in water. Multiply the mass attenuation coefficient with the mass density to obtain the linear attenuation coefficient. When photons are traveling in a particular direction through matter, their number will gradually decrease, since some of them will interact with an electron and

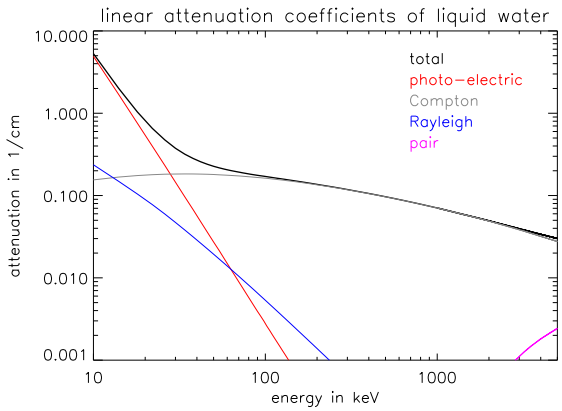

Figure 3.3: *Photon attenuation in water as a function of photon energy*

Figure 3.4: *Positron emitting point source in a non-uniform attenuator.*

get absorbed or deviated into another direction. By definition, the fraction that is eliminated over distance ds equals $\mu(s)N(s)$:

$$-dN(s) = \mu(s)N(s). \tag{3.2}$$

If initially $N(a)$ photons are emitted in point $s = a$ along the s-axis, the number of photons $N(d)$ we expect to arrive in the detector at position $s = d$ is obtained by integrating (3.2):

$$N(d) = N(a)\ e^{-\int_a^d \mu(s)ds}. \tag{3.3}$$

Obviously, the attenuation of a photon depends on where it has been emitted.

For positron emission, a pair of photons need to be detected. Since the fate of both photons is independent, the detection probabilities must be multiplied. Assume that one detector is positioned in $s = d_1$, the second one in $s = d_2$, and a point source in $s = a$, somewhere between the two detectors. Assume further that during a measurement, $N(a)$ photon pairs were emitted along the s-axis (fig.3.4). The number of detected pairs then is:

$$N(d_1, d_2) = N(a)e^{-\int_{d_1}^{a} \mu(s)ds} e^{-\int_a^{d_2} \mu(s)ds} = N(a)e^{-\int_{d_1}^{d_2} \mu(s)ds}. \tag{3.4}$$

Equation (3.4) shows that for PET, the effect of attenuation is independent of the position along the line of detection.

Photon-electron interactions will be more likely if there are more electrons per unit length. So dense materials (with lots of electrons per atom) have a high linear attenuation coefficient.

Chapter 4

Data acquisition

In nuclear medicine, photon detection hardware differs considerable from that used in CT. In CT, large numbers of photons must be acquired in a very short measurement. In emission tomography, a very small number of photons is acquired over a long time period. Consequently, emission tomography systems are optimized for sensitivity.

Photo-electric absorption is the preferred interaction in the detector, since it results in absorption of all the energy of the incoming photon. Therefore the detector material must have a high atomic number (the atomic number Z is the number of electrons in the atom). Since interaction probability decreases with increasing energy, a higher Z is needed for higher energies.

In single photon imaging, ^{99m}Tc is the tracer that is mostly used. It has an energy of 140 keV and the gamma camera performance is often optimized for this energy. Obviously, PET-cameras have to be optimized for 511 keV.

4.1 Detecting the photon

Different detector types exist, but the current standard, both in SPECT and PET, is the scintillation detector. The following sections describe the scintillation crystal, the photomultiplier tube and how crystal and tubes can be combined to make a two-dimensional gamma detector. After that, some alternative designs and new developments will be mentioned.

4.1.1 Scintillation crystal

A scintillation crystal is a remarkable material that stops the incoming photon and in doing so produces a flash of visible light, a scintillation. So the problem of detecting the incoming photon is now transformed in the easier problem of detecting the light flash.

The scintillation stops the photon via photo-electric absorption or via multiple scatter events. The resulting photo-electron travels through the crystal, distributing its kinetic energy over a few thousands other electrons in multiple collisions. As a result, there will be a few thousands of a electrons in an excited state. After a short time, these electrons will release their energy in the form of a photon of a few eV. These secondary photons are visible to the human eye (they are usually blue): this is the scintillation.

The exact wavelength (or color) of the light flash depends on the crystal, but *not* on the energy of the incoming high-energy photon. If a photon with a higher energy enters the

Table 4.1: Characteristics of a few scintillation crystals.

	NaI(Tl)	BGO	LSO
Photons per keV	40	5 . . . 8	20 . . . 30
Scintillation decay time [ns]	230	300	40
Linear atten. coeff. [/cm] (at 511 keV)	0.34	0.95	0.87
Wave length [nm]	410	480	420

crystal, it will send more electrons to an higher energy level. Each of these electrons will then produce a single scintillation-photon, always with the same color. So if we want to have an idea about the energy of the incoming photon, we need to look at the number of scintillation photons (the intensity of the flash), and not at their energy (the color of the flash).

Many scintillators exist, and quite some research on new scintillators is still going on. The crystals that are mostly used today are NaI(Tl) for single photons (140 keV) in gamma camera and SPECT, and BGO (Bismuth Germanate) for annihilation photon (511 keV) in PET. The important characteristics of the crystals are:

- Transparency. If it is not transparent, we cannot see or detect the light flash.
- Photon yield per incoming keV. More photons per keV is better, since the light flash will be easier to see.
- Scintillation time. This is the average time that the electrons remain in the excited state before releasing the scintillation photon. Shorter is better, since we want to be ready for the next photon.
- Attenuation coefficient for the high energy photon. We want to stop the photon, so a higher attenuation coefficient is better. Denser materials tend to stop better.
- Wave length of the scintillation light. Some wave lengths are easier to detect than others.
- Ease of use: some crystals are very hygroscopic: a bit of water destroys them. Others can only be produced at extremely high temperatures. And so on.

The scintillation photons are emitted to carry away the energy set free when an electron returns to a lower energy level. Consequently, these photons contain just the right amount of energy to drive an electron from a lower level to the higher one. In pure NaI this happens all the time, so the photons are reabsorbed and the scintillation light never leaves the crystal. To avoid this, the NaI crystal is doped with a bit of Tl (Thallium). This disturbs the lattice and creates additional energy levels. Electrons returning to the ground state via these levels emit photons that are not reabsorbed. NaI is hygroscopic (similar as NaCl (salt)), so it must be protected against water. LSO (Luthetium ortho-silicate) is a relatively new material. It has very nice features, but is difficult to process (formed at very high temperature), and it is slightly radioactive. Table 4.1 lists some characteristics of three scintillation crystals.

4.1.2 Photo Multiplier Tube

The scintillation crystal transforms the incoming photon into a light flash. That flash must be detected. In particular, we want to know *where* it occurred, *when* it occurred and *how*

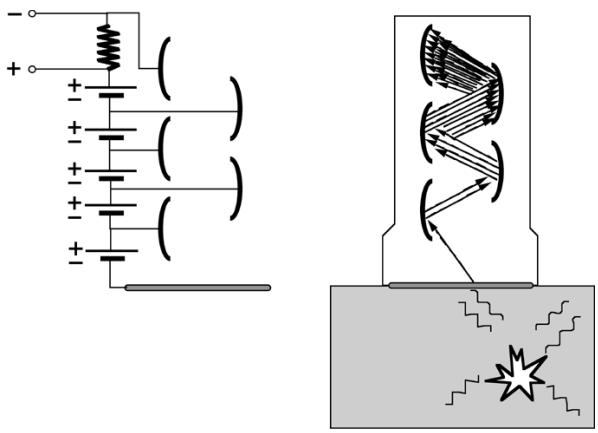

Figure 4.1: *Photomultiplier. Left: the electrical scheme. Right: scintillation photons from the crystal initiate an electric current to the dynode, which is amplified in subsequent stages.*

intense it was (to compute the energy of the incoming photon). All this is usually obtained with photomultiplier tubes (PMT).

The PMT's are glued onto the crystal, in order to detect the incoming scintillation photons. The PMT consists of multiple dynodes, the first of which (the photocathode) is in optical contact with the crystal. High negative voltage (in the order of 100 V) between the dynodes makes the electrons want to jump from one to the other, but they do not have enough energy to cross the gap in between (fig 4.1). This energy is provided by the scintillation photon. The electron acquiring the threshold energy will be accelerated by the electrical field, and activate multiple electrons in the next dynode. After a few steps, a measurable voltage is created which is digitized with an analog-to-digital converter. There are usually about 10 to 12 dynodes, and each step amplifies the signal with a factor $3 \ldots 6$, so amplification can be in the order of one million.

The response time of a PMT is very short (a few ns) compared to the scintillation decay time.

4.1.3 Two-dimensional gamma detector

There are two ways to build a two-dimensional gamma detector based on scintillation. One way is to use a single very large crystal and connect multiple PMT's to it. The other way is to combine small crystals in a large matrix.

4.1.3.1 Single crystal detector

This is the standard design for single photon detectors with NaI(Tl). One side of a single large crystal (e.g. 50 cm x 40 cm, 1 cm thick) is covered completely with PMT's ($\ldots 50 \ldots$ PMTs). The other side is covered with a layer acting as a mirror for the scintillation photons, so that as many as possible are collected in the PMT's.

In principle, all PMT's contribute to the detection of a single scintillation event. As mentioned before, the position, the time and the energy ($\sim$ amount of scintillation photons) must be computed from the PMT-outputs.

- **Position:** (x, y)
  The x-position is computed as
  $$x = \frac{\sum_i x_i S_i}{\sum_i S_i}, \tag{4.1}$$
  where i is the PMT-index, x_i the x-position of the PMT and S_i the integral of the PMT output over the scintillation duration. The y-position is computed similarly. Expression (4.1) is not very accurate and needs correction for systematic errors (linearity correction). The reason is that the response of the PMT's does not vary nicely with the distance to the pulse. In addition, each PMT behaves a bit differently. This will be discussed later in this chapter.

- **Energy:** E
  The energy is computed as
  $$E = c_E \sum_i S_i \tag{4.2}$$
  where c_E is a coefficient converting voltage (integrated PMT output) to energy. The "constant" c_E is not really a constant: it varies slightly with the energy of the high energy photon. Moreover, it depends also on the position of the scintillation, because of the complex and individual behavior of the PMT's. Compensation of all these effects is called "energy correction", and will be discussed below.

- **Time:** t
  In positron emission tomography, we must detect pairs of photons which have been produced simultaneously during positron-electron annihilation. Consequently, the time of the scintillation must be computed as accurately as possible, so that we can separate truly simultaneous events from events that happen to occur almost simultaneously.

  The scintillation has a finite duration, depending on the scintillator (see table 4.1). The scintillation duration is characterized by the the decay time τ, assuming that the number of scintillation photons decreases as $\exp(-t/\tau)$. The decay time ranges from about 40 to several 100 ns. This seems short, but since a photon travels about 1 meter in only 3 ns, we want the time resolution to be in the order of a few ns (e.g. to suppress the random coincidence rate in PET, as will be explained in section 4.2.2.3). To assign such a precise time to a relatively slow event, the electronics computes the time at which a predefined fraction of the scintillation light has been collected.

4.1.3.2 Multi-crystal detector matrix

Instead of using a single large crystal, a large detector area can be obtained by combining many small crystals in a two-dimensional matrix. The crystals should be optically separated (such that scintillation photons are mirrored), to make sure that the scintillation light produced in one crystal stays in that crystal. In the extreme case, each crystal has its own photomultiplier. Computation of position is trivial: it is the coordinate of the crystal in the matrix (no attempt is made to obtain sub-crystal accuracy). Energy is proportional to the total output of the PMT coupled to the crystal. In such a design, the top of the crystal is usually a square of 3 to 5 mm, and the crystal is a cm (140 keV) or a few cm (511 keV) thick. Consequently, the spatial resolution is about 4 mm, as in the single crystal case (see section 4.1.4).

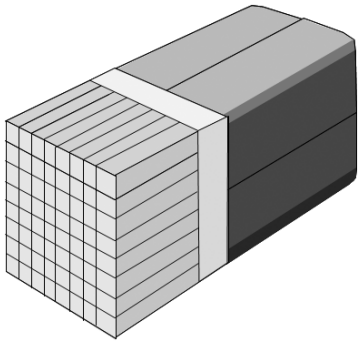

Figure 4.2: *Multi-crystal module. The outputs of four photomultipliers are used to compute in which of the 64 crystals the scintillation occurred.*

Photomultipliers are expensive, so the manufacturing cost can be reduced by a more efficient use of the PMT's. Figure 4.2 shows a typical design, where a matrix of 64 crystals is connected to only four multipliers. The light is guided towards the PMT's in such a way that the PMT amplitude changes monotonically with distance to the scintillating crystal. Thus, the relative amplitudes of the four PMT-outputs allow to compute in which crystal the scintillation occurred.

In a single crystal design, all PMT's contribute to the detection of a single scintillation. If two photons happen to hit the crystal simultaneously, both scintillations will be combined in the calculations and the resulting energy and position will be wrong! Thus, the maximum count rate is limited by the decay time of the scintillation event. In a multi-crystal design, many modules can work in parallel, so count rates can be much higher than in the single crystal design. Count rates tend to be much higher in PET than in SPECT or planar single photon imaging (see below). That is why most PET-cameras are using the multicrystal design, and gamma cameras are mostly single crystal detectors. However, single crystal PET systems and multi-crystal gamma cameras exist as well.

4.1.4 Resolution

The coordinates (x, y, E, t) can only be measured with limited precision. They depend on the actual depth where the incoming photon was stopped, on the number of electrons that was excited, on the time the electrons remain in the excited state, on the direction in which each scintillation photon is emitted when the electron returns to a lower energy state and on the amount of electrons that is activated in the PMT's in each dynode. These are all random processes: if two identical high energy photons enter the crystal at exactly the same position, all the forthcoming events will be different. We can only describe them with probabilities.

So if the photon really enters the crystal at $(\bar{x}, \bar{y}, \bar{E}, \bar{t})$, the actual measurement will produce a random realization (x_i, y_i, E_i, t_i) drawn from a four-dimensional probability distribution (x, y, E, t). We can measure that probability distribution by doing repeated measurements in a well-controlled experimental situation. E.g., we can put a strongly collimated radioactive source in front of the crystal, which sends photons with the same known energy into the crystal at the same known position. This experiment would provide us the distribution for x, y and E. The time resolution can be determined in several ways, e.g. by activating the PMT's with light emitting diodes. The resolution on x and y is often called the *intrinsic*

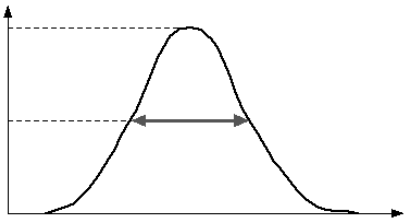

Figure 4.3: *The full width at half max of a probability distribution*

resolution.

If we plot all the measurements in a histogram, we will see the distribution. Usually the distribution is approximately Gaussian, and can be characterized by its standard deviation σ. Often one specifies the full width at half maximum (FWHM) instead (figure 4.3). It is easy to show that for a Gaussian, the FWHM $= 2\sqrt{2\ln 2}\sigma$. This leads to a useful rule of thumb: *anything smaller than the FWHM is lost during the measurement.*

Position resolution of a scintillation detector has a FWHM of 3 to 4 mm. Energy resolution is about 10% FWHM (so 14 keV for a 140 keV tracer) in NaI(Tl) cameras, and about 25% FWHM (about 130 keV at 511 keV) or worse in BGO PET-systems.

4.1.5 Storing the data

To store the data, the information must be digitized. We do not want to lose information, so the round-off errors made during digitization should be small compared to the resolution. Thus, we obtain four digital numbers (x, y, E, t), which are the coordinates of a single detected photon. If all information must be preserved, we can directly store the four numbers in a list (list-mode acquisition). Obviously, this list may become very long. Often, we do not need all this information. The energy is usually simply used to decide whether we will accept or keep the photon, it does not need to be stored (this will be explained in section 4.3). Similarly, we often want to make a single image representing the situation in a certain time interval, so we only need a single two-dimensional image. To store this information efficiently, we prepare a zero two-dimensional matrix, an "empty" image, and increment the value at coordinates x_i, y_i for the i-th accepted photon. The number can be interpreted as a brightness, so we can display the result as an image on the screen. Brightness (or some pseudo-color) corresponds to tracer concentration.

4.1.6 Alternative designs

Although most current gamma cameras and PET systems are based on scintillation crystals combined with photomultipliers, there are many other detection systems. Some of these have very good characteristics, but their acceptance is hampered by the high cost price and, in some cases, by the fact that new algorithms must be developed before their full potential can be exploited. In the following, we only mention two promising technologies. Avalanche photodiodes can replace the PMT, CdZnTe detectors can replace the entire standard detection system.

4.1.6.1 Avalanche photodiodes

A semiconductor is a material in which the large majority of electrons are strongly bound to the atoms, but not all of them. Some electrons have enough energy to move freely about as in a metal. As a result, some of the atoms have lost an electron and have a net positive charge. This phenomenon gives rise to two types of charge carriers that can support electric current: electrons and holes. The former are the freely moving electrons. The latter are the positive vacancies left behind by the electrons. The holes can move, because an electron from a neighboring atom can jump over to fill in the vacancy, creating a similar vacancy in the neighboring atom.

Electric current is possible if freely moving electrons or holes are available. However, if for some reason none of them are available, the material acts as an insulator.

A diode is a simple electronic device with wonderful behavior. It has excellent conductivity in one direction, and extremely poor conductivity in the other. This remarkable feature is obtained by connecting two very different types of semiconductor materials, resulting in a very asymmetrical distribution of charge carriers. When forward voltage is applied, charge carriers are injected in great numbers, leading to very low resistance. When reverse voltage is applied, the electrons and holes are pulled away from the junction, so there is nothing left to support an electric current. Consequently, it behaves as an insulator in this mode.

In an avalanche photodiode (APD), a very high reverse voltage is applied, pulling away all charge carriers. However, a photon hitting the diode may supply enough energy to break the covalent bond of an electron, thus creating an electron-hole pair. This free electron will move rapidly in the electric field, releasing other electrons from their bond. Consequently, the photon creates an avalanche of free electrons, resulting in a measurable current.

APD's can be used to replace the photomultiplier. The advantage is that they are much smaller, but currently they are still more expensive than the PMTs.

4.1.6.2 Solid state detectors

An effect similar to the one described above can be used to directly detect the high energy photon instead of the scintillation photons. For this, a material with high stopping power is required. This is the operating principle of the CdZnTe detector: a strong electric field drives charge carriers created by the high energy photon towards a grid of collectors. Position resolution is now determined by the size of the collectors. Designs with sub-millimeter resolution exist. Energy resolution is excellent (3 %, as opposed to the 10 % in NaI(Tl)). And also the stopping power is good: 122 keV photons have a mean traveling distance of less than 2 mm. Again, the price is a problem.

As will be discussed in the next section, the resolution of a camera is dominated by the collimator acceptance angle. Consequently, the excellent spatial resolution of the CdZnTe detector is completely wasted in the traditional design. Some researchers are studying alternative camera designs and accompanying algorithms to fully exploit the excellent performance of this type of detectors.

4.2 Collimation

At this point, we know how the position and the energy of a photon impinging on the detector can be determined. Now we need a way to ensure that the impinging photons will produce

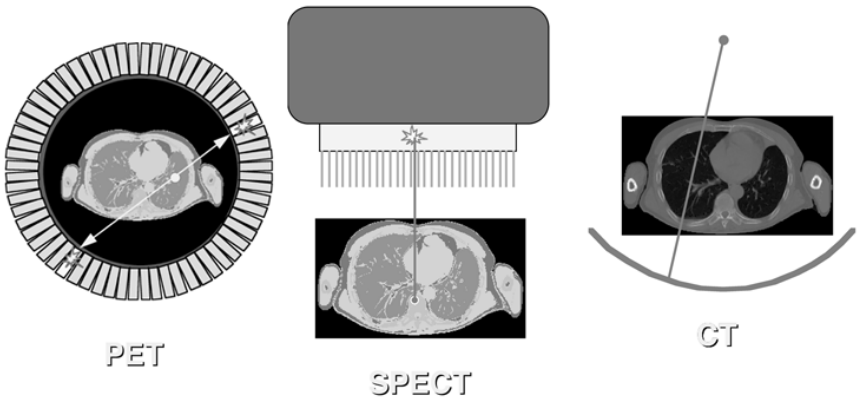

Figure 4.4: *Collimation in the PET camera, the gamma camera (SPECT) and the CT camera.*

an image. In photography, a lens is used for that purpose. But there are no lenses to focus the high energy photons used in nuclear medicine. So we must fall back on a more primitive approach: collimation. Collimation is the method used to make the detector "see" along straight lines. Different tomographic systems use different collimation strategies, as show in figure 4.4.

In the CT-camera, collimation is straightforward: there is only one transmission source, so any photon arriving in the detector has traveled along the straight line connecting source and detector. In the PET-camera, a similar collimation is obtained: if two photons are detected simultaneously, we can assume they have originated from the same annihilation, which must have occurred somewhere on the line connecting the two detectors. But for the gamma camera there is a problem: only one photon is detected, and there is no simple way to find out where it came from. To solve that problem, mechanical collimation is used. A mechanical collimator is essentially a thick sieve with long narrow holes separated by thin septa. The septa are made from a material with strong attenuation (usually lead), so photons hitting a septum will be eliminated, mostly by photo-electric interaction. Only photons traveling along lines parallel to the collimator holes will reach the detector. So instead of computing the trajectory of a detected photons, we eliminate all but one trajectory, so that we know the trajectory even before the photon was detected. Obviously, this approach reduces the sensitivity of the detector, since many photons will end up in the septa. This is why a PET-system acquires more photons per second than a gamma camera, for the same activity in the field of view.

Most often, the parallel hole collimator is used, but for particular applications other collimators are used as well. Figure 4.5 shows the parallel hole collimator (all lines parallel), the fan beam collimator (lines parallel in one plane, focused in the other), the cone beam collimator (all lines focused in a single point) and the pin hole collimator (single focus point, but in contrast with other collimators, the focus is placed *before* the object).

Collimation ensures that tomographic systems collect information about lines. In CT, detected photons provide information about the total attenuation along the line. In gamma cameras and PET cameras, the number of detected photons provide information about the total radioactivity along the line (but of course, this information is affected by the attenuation as well). Consequently, the acquired two-dimensional image can be regarded as a set of line integrals of a three dimensional distribution. Such images are usually called *projections*.

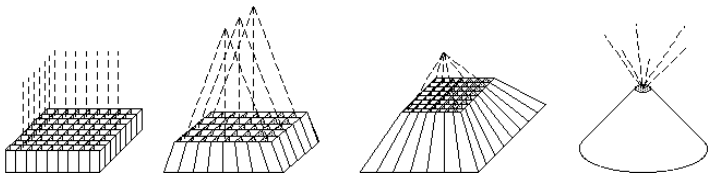

Figure 4.5: *Parallel hole, fan beam, cone beam and pin hole collimators.*

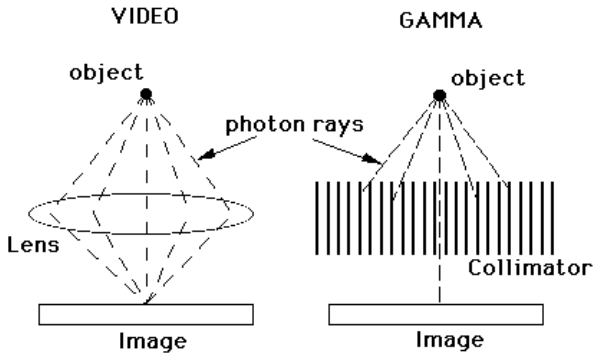

Figure 4.6: *Focusing by a lens compared to ray selection by a parallel hole collimator*

4.2.1 Mechanical collimation in the gamma camera

Figure 4.6 shows that the sensitivity obtained with a mechanical collimation is very poor compared to that obtained with a lens. The mechanical collimator has a dominating effect on the resolution and sensitivity of the gamma camera. That is why a more detailed analysis of the collimator point spread function is in order.

Fig. 4.7 shows the cross section of a parallel hole collimator with septa length T and septa spacing a. At a distance of H, a radioactive point source is located. If the collimator would really absorb all photons except those propagating exactly perpendicular to the detector, then the image would be a point. However, fig. 4.7 shows that photons propagating along slightly inclined lines may also reach the detector. Therefore the image will not be a point. Because the image of a point is characteristic for the system, such images are often studied. By definition, the image of a point is the *point spread function* (PSF) of the imaging system.

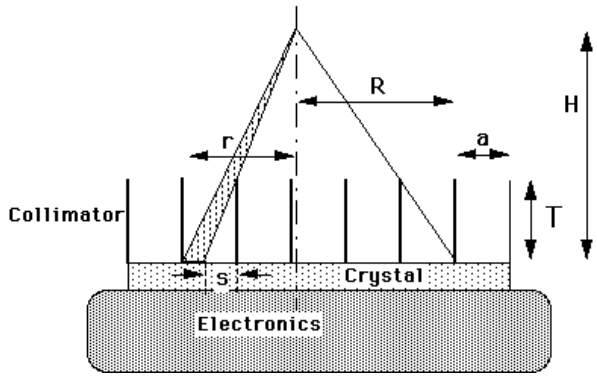

Figure 4.7: *Parallel hole collimator.*

Assume that the thickness of the septa can be ignored, that H is large compared to the length of the septa T, and that T is large compared to a. We will also ignore septal penetration (i.e. gamma-photons traveling through the septa instead of getting absorbed). The light (e.g. 140 keV photons) emitted by the source makes the septa cast a shadow on the detector surface. In fact, most of the detector is in that shadow, except for a small region facing the source. We will regard the center of that region, immediately under the source, as the origin of the detector plane. The length of the shadow s cast by a septum on the detector is then

$$s = r\frac{T}{H} \tag{4.3}$$

where r is the distance to the origin. The fraction of the detector element that is actually detecting photons is

$$f = \frac{a-s}{a} = 1 - \frac{rT}{aH} \tag{4.4}$$

This expression holds for r between 0 and R, where R is the position at which the shadow completely covers the detector so that $f = 0$. The expression gives the fraction of the photons detected at r, relative to the number that would be detected at the same position if there was no collimator. That number is easy to compute. The source is emitting photons uniformly, so the number of photons per solid angle is constant. Stated otherwise, if we would put a point source in the center of a spherical uncollimated detector with radius H, the detector surface would be uniformly irradiated. The total area of the sphere is $4\pi H^2$, so the sensitivity per unit area is $1/(4\pi H^2)$. Multiplying with the collimator sensitivity produces the point spread function of the collimated detector at distance H:

$$\mathrm{PSF}(r) = (1 - \frac{rT}{aH})\frac{1}{4\pi H^2} \tag{4.5}$$

In this expression, we have ignored the fact that for a flat detector, the distance to the point increases with r. The approximation is fair if r is much smaller than H. Consequently, the PSF has a triangular profile, as illustrated in figure 4.8.

To calculate (approximately) the total collimator sensitivity, expression (4.5) must be integrated over the region where $\mathrm{PSF}(r)$ is non-zero. We assume circular symmetry and use integration instead of the summation (actually required by the discrete nature of the problem). Integration assumes that a and T are infinitely small. Since we use only the ratio, this poses no problems.

$$\begin{aligned} \text{sens} &= \frac{1}{4\pi H^2}\int_0^R (1 - \frac{rT}{aH}) 2\pi r dr && (4.6)\\ &= \frac{1}{4\pi H^2}\frac{\pi R^2}{3} = \frac{1}{12}\left(\frac{a}{T}\right)^2 && (4.7)\\ R &= \frac{aH}{T} && (4.8) \end{aligned}$$

With a septa length of 2 cm and septa spacing of 1 mm, about 1 in 5000 photons is passing through the collimator, according to equation (4.7).

It is easy to show that (4.8) is also equal to the FWHM of the PSF: you can verify by computing the FWHM from (4.5). So the FWHM of the spatial resolution *increases linearly with the distance to the collimator!* Therefore, the collimator must always be as close as possible to the patient, and failure to do so leads to important image degradation!

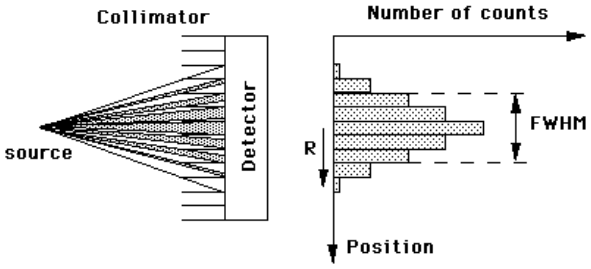

Figure 4.8: *Point spread function of the collimator. The number of detected photons at a point in the detector decreases linearly with the distance r to the central point.*

To improve the resolution, the ratio a/T must be decreased. However, the sensitivity is proportional to the square of a/T. As a result, the collimator must be a compromise between high sensitivity and low FHWM.

The PSF is an important characteristic of the linear system, which allows us to predict how the system will react to any input. See appendix 10.2 if you are not familiar with the concept of PSF and convolution integrals.

In section 4.1.4 we have seen that the *intrinsic resolution* of a typical scintillation detector has a FWHM of about 4 mm. Now, we have seen that the collimator contributes significantly to the FWHM, but in the derivation, we have ignored the intrinsic resolution. Obviously, they have to be combined to obtain realistic results, so we must "superimpose" the two random processes. The probability distribution of each random process is described as a PSF. To compute the overall probability distribution, the second PSF must be "applied" to every point of the first one. Mathematically, this means that the two PSFs must be convolved. If we make the reasonable assumption that the PSFs are Gaussian, convolution becomes very simple, as shown in appendix 10.3: the result is again a Gaussian, and its variance equals the sum of variances of the contributing Gaussians.

At distances larger than a few cm, the collimator PSF dominates the spatial resolution. The PSF increases in the order of half a cm per 10 cm distance, but there is a wide range: different types of collimator exist, either focusing on resolution or on sensitivity.

4.2.2 Electronic and mechanical collimation in the PET camera

4.2.2.1 Electronic collimation: coincidence detection

During positron annihilation, two photons are emitted in opposite directions. If both are detected, we know that the annihilation occurred somewhere on the line connecting the two detectors. So in contrast to SPECT, no mechanical collimator is needed. However, we need fast electronics, since the only way to test if two detected photons could belong together, is by checking if they have been emitted simultaneously.

We first consider a single detector pair as shown in figure 4.9, and study its characteristics by computing its response to a point source. The detector has a square surface of size d, and the distance between the detectors is R. We chose the origin in the point of symmetry, and see how the response changes if we change the position of the point source. Keep in mind that an event is only valid if both photons of the photon pair are detected.

Assume that the point source is moved along the r-axis: $x = 0$. Then the limiting detector is the far one: if one photon reaches the far detector, the other photon will surely

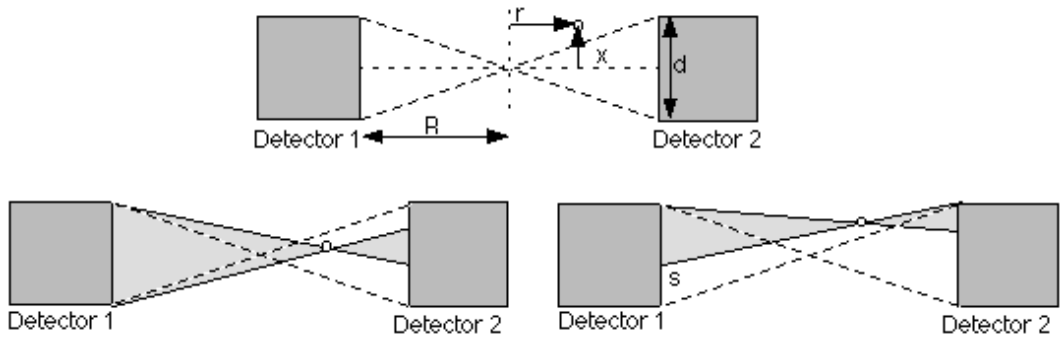

Figure 4.9: *A point source positioned in the field of view of a pair of detectors*

.

reach the other detector. The number of detected photon pairs is proportional to the solid angle occupied by the two detectors:

$$\mathrm{PSF}(x=0,r) = \frac{2d^2}{4\pi(R+r)^2} = \frac{d^2}{2\pi(R+r)^2} \tag{4.9}$$

This is an approximation: we assume that the far detector surface coincides with the surface of a sphere of radius $R+r$. It does not, because it is flat. However, because d is very small compared to R, the approximation is very good.

We now move the point source vertically over a distance x. As long as the point is between the dashed lines in figure 4.9, nothing changes (that is, if we ignore the small decrease in solid angle due to the fact that the detector surface is now slightly tilted when viewed from the point source). However, when the point crosses the dashed line, the far detector is no longer fully irradiated. A part of the detector surface no longer contributes: if a photon reaches that part, the other photon will miss the other detector. The height s is computed from simple geometry (figure 4.9):

$$s(x) = \left(|x| - \frac{dr}{2R}\right)\frac{2R}{R-r} \tag{4.10}$$

The first factor is the distance between the point and the dashed line, the second factor is the magnification due to projection of that distance to the far detector. This equation is only valid when $dr/(2R) \leq x \leq d/2$, $s=0$ if x is smaller and $s=d$ if x is larger. Knowing the active detector area, we can immediately compute the corresponding solid angle to obtain the PSF:

$$\mathrm{PSF}(x,r) = \frac{(d-s(x))d}{2\pi(R+r)^2} \tag{4.11}$$

We can move the point also in the third dimension, which we will call y. The effect is identical to that of changing x, so we obtain:

$$\mathrm{PSF}(x,y,r) = \frac{(d-s(x))(d-s(y))}{2\pi(R+r)^2} \tag{4.12}$$

From these equations it follows that the PSF is triangular in the center, rectangular close to the detectors and trapezoidal in between. Introducing that third dimension, we obtain for

Table 4.2: Kinetic energy, maximum and mean path length of the positron for the most common PET isotopes.

Isotope	Kinetic energy [MeV]	Maximum path [mm]	Mean path [mm]
^{11}C	0.96	3.9	1.1
^{13}N	1.19	5.1	1.5
^{15}O	1.72	8.0	2.5
^{18}F	0.64	2.4	0.6
^{68}Ga	1.90	8.9	2.9
^{82}Rb	3.35	17	5.9

the PSF in the center and close to the detector respectively:

$$\mathrm{PSF}(x,y,0) = \frac{(d-2|x|)(d-2|y|)}{2\pi R^2} \tag{4.13}$$

$$\mathrm{PSF}(x,y,R) = \frac{d^2}{8\pi R^2} \tag{4.14}$$

We can compute the average sensitivity within the field of view by integrating the PSF over the detector area $x \in [-d/2, d/2], y \in [-d/2, d/2]$ and divide by d^2. It is simple to verify that in both cases we obtain:

$$\mathrm{sens} = \frac{d^2}{8\pi R^2} \tag{4.15}$$

showing that sensitivity is independent of position in the detection plane, if the object is large compared to d.

Finally, we can estimate the sensitivity of a complete PET system consisting of a detector ring with thickness d and radius R, assuming that we have a positron emitting source near the center of the ring. We assume that the ring is in the (y,r)-plane, so the x-axis is the symmetry axis of the detector ring. One approach is to divide the detector area by the area of a sphere with radius R, as we have done before. This yields:

$$\mathrm{PETsens}_{x=0} = \frac{2\pi dR}{4\pi R^2} = \frac{d}{2R} \tag{4.16}$$

This is the maximum sensitivity, obtained for $x = 0$. The average over $= -d/2 \ldots d/2$ is half that value, since in the center of the PET, the sensitivity varies linearly between the maximum and zero:

$$\mathrm{PETsens} = \frac{d}{4R}. \tag{4.17}$$

4.2.2.2 Resolution of coincidence detection

Until now, we have always assumed that annihilation takes place very close to the emission point, and that the photons are emitted in exactly opposite directions. In reality, there are small deviations, which limit the theoretical resolution of PET.

As shown in table 4.2, the path depends on the kinetic energy of the positron. The positron can only annihilate if its kinetic energy is sufficiently low, so if it is emitted at high speed,

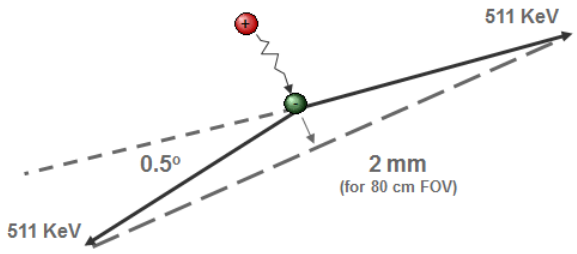

Figure 4.10: *The distance traveled by the positron and the deviation from 180 degrees in the angle between the photon paths limit the best achievable resolution in PET*

.

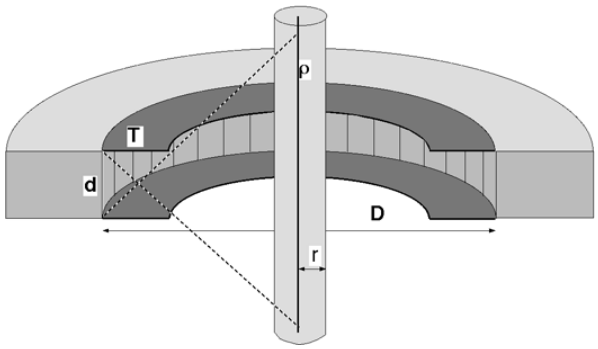

Figure 4.11: *A PET ring detector (cut in half) with a cylindrical homogeneous object, containing a radioactive wire near the symmetry axis*

.

it will travel far. During its trajectory, it dissipates its energy in collisions with surrounding electrons.

Momentum is preserved during annihilation, so the momentum of both photons must equal the momentum of the positron and the electron. This momentum is usually not zero, so the photons cannot be emitted in exactly opposite directions (the momentum of a photon is a vector with amplitude h/λ and pointing in the direction of propagation). As shown in figure 4.10, there is a deviation of about 0.3°, which corresponds to a 2.8 mm offset for a detector ring with 1 m diameter.

4.2.2.3 Mechanical collimation: inter-plane septa

In the previous section we have seen why a PET-camera does not need a collimator. We also know that the PET-camera can only measure photon pairs originating somewhere within the detection plane. However, photons coming from radioactivity outside that plane could also reach the detectors, and produce undesired scintillations, which provide no information, or even worse, wrong information. Therefor, it is useful to shield the detector ring against photons coming from outside the field of view. This is done with so-called septa, as shown in figure 4.11. The septa provide some collimation in the direction perpendicular to the detection plane, but there is no collimation within the plane. The septa reduce the number of single photons, scattered photons, random coincidences and triple (or more) coincidences.

Before these events are described, we have to better define what is meant with a coinci-

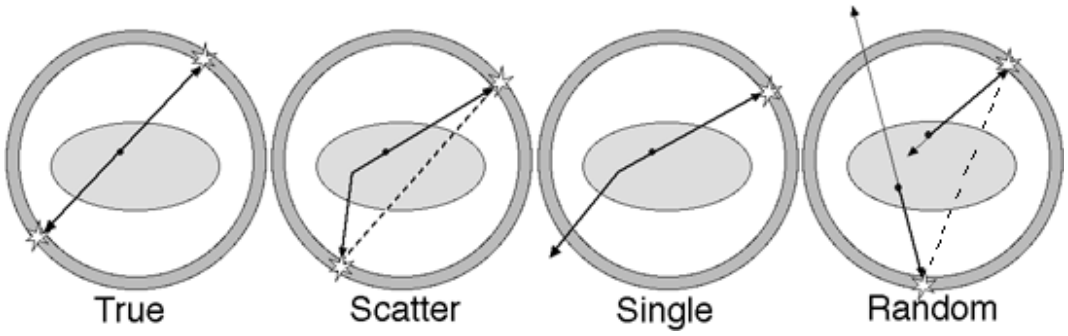

Figure 4.12: *True coincidence, scattered coincidence, single event, random coincidence*

.

dence. Two photons are detected *simultaneously* if the difference in detection times is lower than a predefined threshold, the "time window" of the PET-system. The time window cannot be arbitrarily short for the following reasons:

- the time resolution of the electronics is limited: electricity is never faster than light, and it is delayed in every electronical component.
- The diameter of the PET-camera is typically about 1 m for a whole body system, and about half that for a brain system. Light travels about 1 m in 3 ns. In theory, one could attempt to derive position information from the difference in arrival time. This is called TOF-PET (Time Of Flight). However, for several technical reasons, current clinical systems do not attempt to exploit time of flight. Consequently, the time window should not be below 3 ns for a 1 m diameter system.
- The time resolution is limited. We do not want to reject true coincidences because of errors in timing computation.

For these reasons, current systems use a time window of about 10 ns.

Figure 4.12 shows the difference between a true coincidence, which provides valuable information, and several other events which are disturbing or even misleading. A single event provides no information and is harmless. But if two single events are simultaneous (random coincidence), they are indistinguishable from a true coincidence. If one (or both) of the photons is scattered, the coincidence provides wrong information: the decaying atom is not located on the line connecting the two detectors. Finally, a single event may be simultaneous with a true coincidence. Since there is no way to find out which of the three photons is the single one, the entire event is ignored and information is lost.

To see how the septa dimensions influence the relative contribution of the various possible events, some simple computations can be carried out using the geometry shown in figure 4.11. Assume that we have a cylindrical object in the PET-camera. In the center of the cylinder, there is a radioactive wire, with activity ρ per unit length. The cylinder has radius r and attenuation coefficient μ. This object represents an idealized patient: there is activity, attenuation and Compton scatter in the attenuating object as in a patient; the unrealistic choice of dimensions makes the computations easier. The PET-camera diameter is D, the detector size is d and the septa have length T. The duration of the time window is τ. Finally, we assume that the detector efficiency is ϵ, which means that if a photon reaches the detector,

it has a chance of ϵ to produce a valid scintillation, and a chance of $1-\epsilon$ to remain undetected.

To compute the number of **true coincidences** per unit of time, we must keep in mind that both photons must be detected, but that their paths are not independent. To take into account the influence of the time window, it helps to consider first a short time interval equal to the time window τ. After that, we convert the result to counts per unit of time by dividing by τ. Why we do this will become clear when dealing with randoms. We proceed as follows:

- The activity in the field of view is ρd.
- Both photons can be attenuated, the attenuated activity is $\rho d a^2$, with $a = \exp(-\mu r)$.
- As we have derived above, the geometrical sensitivity is $d/(2D)$.
- Both photons must be detected if they hit the detectors, so the effective sensitivity is $\epsilon^2 d/(2D)$.
- The number of counts is proportional to the scan time τ.
- To compute the number of counts per time unit, we multiply with $1/\tau$.

We only care about the influence of the design parameters, so we ignore all constants. This leads to:

$$\text{trues} \sim \rho a^2 \epsilon^2 \frac{d^2}{D} \tag{4.18}$$

For **scatters** the situation is very similar. The only difference is that the combined photon path is a broken line. This has two consequences. First, the field of view is larger than for the trues. Second, the path of the second photon is (nearly) independent of that of the first one.

We will assume that only one of the photons undergoes a single Compton scatter event, so the path contains only a single break point. The corresponding field of view is indicated with the dashed line in figure 4.11. The length of the wire in the field of view is then $d(D-T)/T$, but since T is much smaller than D we approximate it as dD/T.

Each of the photons goes its own way. This will only lead to detection if both happen to end up on a detector and are detected. For each of them, that probability is proportional to $\epsilon d/D$, so for the pair the detection probability is $(\epsilon d/D)^2$. For the trues this factor was not squared, because detection of one photon guarantees detection of the other!

Attenuation of scattered photons is complex, since they have lower energy and travel along oblique lines. However, if we assume that the septa are not too short and that the energy resolution is not too bad, we can ignore all that without making too dramatic an error. (We only want to see the major dependencies, so we can live with moderately dramatic errors.) Consequently, we have for the scatters count rate

$$\text{scatters} \sim \rho a^2 \epsilon^2 \frac{d^3}{DT} \tag{4.19}$$

A **single**, non-scattered photon travels along a straight line, so the singles field of view is the same as that of the scatters, and the contributing activity is $\rho d D/T$. The attenuation is a. The effective sensitivity is proportional to $\epsilon d/D$. The number of singles in a time τ is

proportional to τ. Finally, we divide by τ to obtain the number of single events per unit of time. This yields:

$$\text{singles} \sim \rho a \epsilon \frac{d^2}{T} \tag{4.20}$$

A **random coincidence** consists of two singles, arriving *simultaneously*. So if we study a short time interval τ equal to the time window, we must simply multiply the probabilities of the two singles, since they are entirely independent of each other. Afterwards, we multiply with $1/\tau$ to compute the number of counts per time unit. So we obtain:

$$\text{randoms} \sim \rho^2 a^2 \epsilon^2 \frac{d^4}{T^2} \tau \tag{4.21}$$

Similarly, we can compute the probability of a "**triple coincidence**", resulting from simultaneous detection of a true coincidence and a single event. This produces:

$$\text{true} + \text{single} \sim \rho^2 a^3 \epsilon^3 \frac{d^4}{TD} \tau \tag{4.22}$$

As you can see, the trues count rate is not affected by either τ nor T, in contrast to the unwanted events. Consequently, we want τ to be as short as possible to reduce the randoms and triples count rate. We have seen that for current systems this is about 10 ns. Similarly, we want T to be as large as possible to reduce all unwanted events. Obviously, we need to leave some room for the patient.

4.2.2.4 2D and 3D PET

In the previous paragraphs we have studied a single ring of PET detectors, and we have shown that shielding the detectors with septa reduces the scatter and random coincidence rate, without affecting the true coincidence rate. In current clinical systems, multiple rings are combined in a single device, as shown in figure 4.13. Many of these systems can be operated in two modes. In 2D-mode, the rings are separated by septa. In 3D-mode, the septa are retracted, only the shielding at the edges of the axial field of view remain.

In 2D mode, there are two types of planes: direct planes, located in the center of the ring, and cross-planes, located between two neighboring rings. The direct planes correspond to what we have seen above for a single ring system. A photon pair belongs to a cross plane if one photon is detected in ring i and the other one in ring $i+1$. The small inclination of the projection line can be ignored, and such coincidences are processed as if they were acquired by an imaginary detector ring located at $i+0.5$. This improves axial sampling. It turns out that the cross-planes are in fact superior to the direct planes: near the center the axial resolution is slightly better because of the shadow cast by the septa on the contributing detectors, and the sensitivity is nearly twice that of direct planes because of the larger solid angle.

According to the Nyquist criterion, the sampling distance should not be longer than half the period of the highest special frequency component in the acquired signal (you need at least two samples per period to represent a sine correctly). Without the cross-planes, the criterion would be violated. So the cross-planes are not only useful to increase sensitivity, the are needed to avoid aliasing as well.

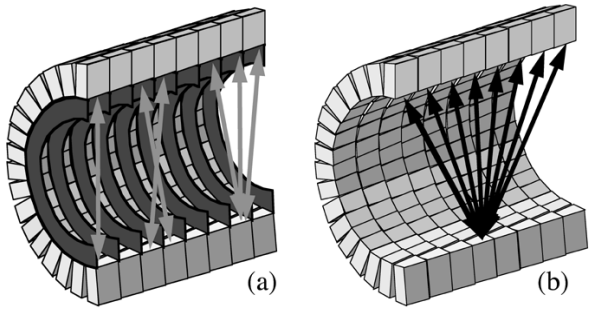

Figure 4.13: *Positron emission tomograph (cut in half). When septa are in the field of view (a), the camera can be regarded as a series of separate 2D systems. Coincidences along oblique projection lines between neighboring rings can be treated as parallel projection lines from an intermediate plane. This doubles axial sampling: 15 planes are reconstructed from 8 rings. Retracting the septa (b), increases the number of projection lines and hence the sensitivity of the system, but fully 3D reconstruction is required.*

In 3D mode, the septa are retracted, except for the first and last one. The detection is no longer restricted to parallel planes, photon pairs traveling along oblique lines are now accepted as well. In chapter 5 we will see that this has a strong impact on image reconstruction.

This has no effect on spatial resolution, since that is determined by the detector size. But the impact on sensitivity is dramatic. In fact, we have already computed all the sensitivities in section 4.2.2.3. Those expressions are still valid for 3D PET, if we replace d (the distance between the septa) with Nd, where N is the number of neighboring detector rings. To see the improvement for 3D mode, you have to compare to a concatenation of N independent detector rings (2D mode), which are N times more sensitive for any event than a single ring. So you can see that the sensitivity for trues increases with a factor of N when going from 2D to 3D. However, scatters increase with N^2 and randoms even with N^3. Because the 3D PET is more sensitive, we can decrease the injected dose with a factor of N (preserve the count rate). If we do that, randoms will only increase with N^2. Consequently, the price we pay for increased sensitivity is an even larger increase of disturbing events. So scatter and randoms correction will be more important in 3D PET than in 2D PET. It turns out that in particular scatter poses problems: it was usually ignored in 2D PET, but this is no longer acceptable in 3D PET. Various scatter correction algorithms for 3D PET have been proposed in the literature, and one of those is being used systematically in clinical 3D PET systems.

4.3 Energy windowing

Compton scatter causes photons to be deviated from their original trajectory, and to propagate with reduced energy along a new path. Consequently, photons with reduced energy can reach the detector via a broken line. Such photons are harmful: they provide no useful information and produce an unwanted inhomogeneous background in the acquired images (fig 4.14).

We have seen in section 3.2 that the photon loses more energy for larger scatter angles. The gamma camera (or the PET camera) measures the energy of the photon, so it can reject the photon if its energy is low.

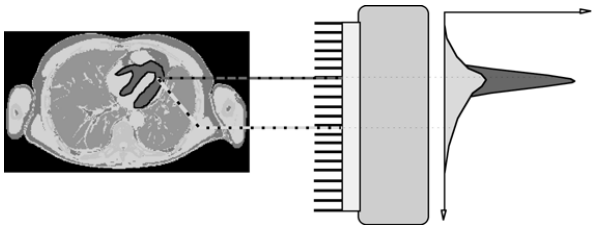

Figure 4.14: *Compton scatter allows photons to reach the camera via a broken line. These photons produce a smooth background in the acquired projection data.*

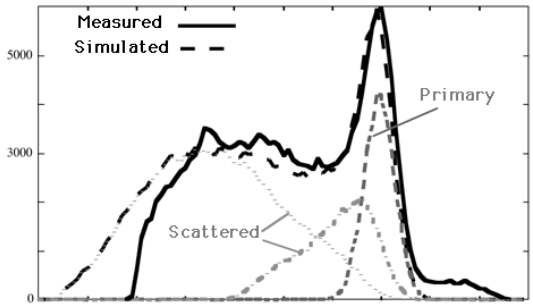

Figure 4.15: *The energy spectrum measured by the gamma camera, with a simulation in overlay. The simulation allows to compute the spectrum of non-scattered (primary) photons, single scatters and multiple scatters. The true primary spectrum is very narrow, the measured one is widened by the limited energy resolution.*

However, if the energy loss during Compton scatter is small or similar relative to the energy resolution of the gamma camera, then it may survive the energy test and get accepted by the camera. Figure 4.15 shows the energy spectrum as measured by the gamma camera. A Monte Carlo simulation was carried out as well, and the resulting spectrum is very similar to the measured one. The simulation allows to compute the contribution of unscattered (or primary) photons, photons that scattered once and photons that suffered multiple scatter events. If the energy resolution were perfect, the primary photon peak would be infinitely narrow and all scattered photons could be rejected. But with a realistic energy resolution the spectra overlap, and acceptance of scattered photons is unavoidable.

Figure 4.15 shows a deviation between simulation and measurement for high energies. In the measurement it happens that two photons arrive simultaneously. If that occurs, the gamma camera adds their energies and averages their position, producing a mispositioned count with high energy. Such photons must be rejected as well. (The simulator deviates also at low energies from the measured spectrum, because the camera has not recorded these low-energy photons.)

To reject as much as possible the unwanted photons, a narrow energy window is centered around the tracer energy peak, and all photons with energy outside that window are rejected. Current gamma camera can use multiple energy windows simultaneously. That allows us to administer two or more different tracers to the patient at the same time, if we make sure that each tracer emits photons at a different energy. The gamma camera separates the photons

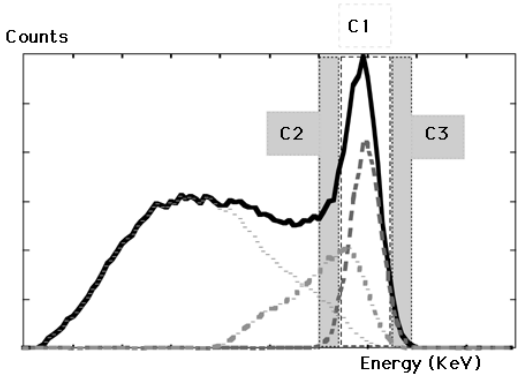

Figure 4.16: *Triple energy window correction. The amount of scattered photons accepted in window C1 is estimated as a fraction of the counts in windows C2 and C3.*

based on their energy and produces a different image for each tracer.

Since not all unwanted photons can be rejected, an additional correction may be required. Figure 4.16 shows a typical correction method based on three energy windows. Window C1 is the window centered on the energy of the primary photons. If we assume that the spectrum of the unwanted photons varies approximately linearly over the window C1, then we can estimate that spectrum by measuring in two additional windows C2 (just below C1) and C3 (just above C1). If all windows had the same size, then the number of unwanted photons in C1 could be estimated as (counts in C2 + counts in C3) / 2. Usually C2 and C3 are chosen narrower than C1, so the correction must be weighted accordingly. In the example of figure 4.16 there are very few counts in C3. However, if we would use an second tracer with higher energy, then C3 would receive scattered photons from that tracer.

4.4 Corrections

Currently, most gamma camera designs are based on a single crystal detector as shown in figure 4.17. To increase the sensitivity, a single gamma camera may have two or three detector heads, enabling simultaneous acquisition of two or three views along different angles simultaneously. Because the detector head represents a significant portion of the cost of the gamma camera, the price increases rapidly with the number of detector heads.

Most PET systems have a multi-crystal design, they consist of multiple rings (fig 4.13) of small detectors. The detectors are usually arranged in modules (fig 4.2) to reduce the number of PMT's.

The performance of the PMT's is not ideal for our purposes, and in addition they show small individual differences in their characteristics. In current gamma cameras and PET cameras the PMT-gain is computer controlled and procedures are available for automated PMT-tuning. But even after tuning, small differences in characteristics are still present. As a result, some corrections are required to ensure that the acquired information is reliable.

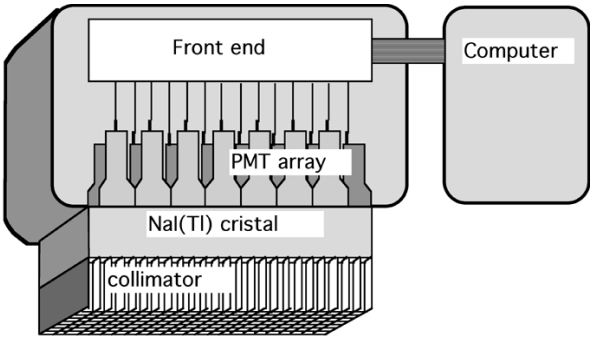

Figure 4.17: *Schematic representation of a gamma camera with a single large scintillation crystal and parallel hole collimator.*

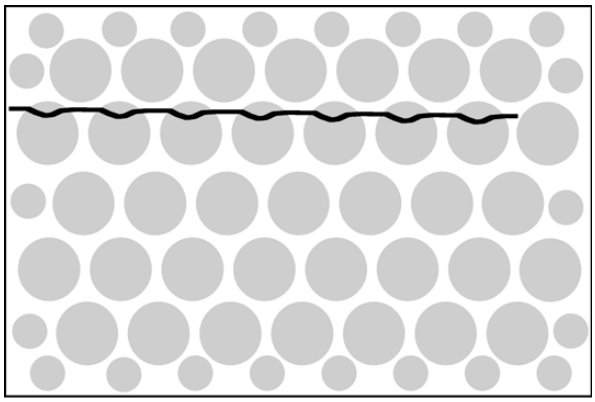

Figure 4.18: *The PMTs (represented as circles) have a non-linear response as a function of position. As a result, the image of a straight radioactive line would be distorted as in this drawing.*

4.4.1 Linearity correction

In section 4.1.3.1 we have seen that the position of a scintillation in a large crystal is computed as the first moment in x and y directions (the "mass" center of the PMT response). This would be exact if the PMT response would vary linearly with position. In reality, the response is not perfectly linear, and as a result, the image is distorted. Figure 4.18 illustrates how the image of a straight radioactive wire can be deformed by this non-linear response. The response is systematic, so it can be measured, and the errors Δ_x and Δ_y can be stored as a function of x and y in a lookup table. Correction is then straightforward:

$$\begin{aligned} x_{\text{corrected}} &= x + \Delta_x(x, y) \\ y_{\text{corrected}} &= y + \Delta_y(x, y) \end{aligned} \tag{4.23}$$

After correction, the image of a straight line will be a straight line.

For a system using multi-crystal detectors, the linearity correction must be applied within the modules, to make sure the correct individual detector is identified. Because we know exactly where each module and each detector is located, no further spatial corrections are required.

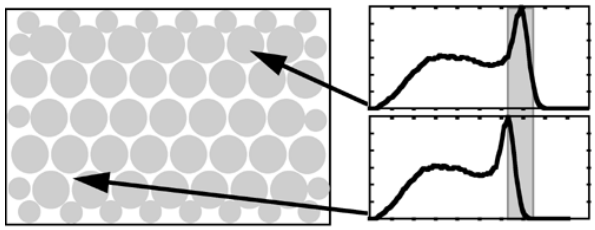

Figure 4.19: *Because the characteristics of the photomultipliers are not identical, the conversion of energy to voltage (and therefore to a number in the computer) may be position dependent.*

Because the system characteristics vary slowly in time, the linearity correction table needs to be measured every now and then. This is typically once a year or after a major intervention (e.g. after PMTs have been replaced).

To measure the linearity, we must make an image of a *phantom*, a well known object. One approach is to use a "dot-phantom": the collimator is replaced with a lead sheet containing a matrix of small holes at regular positions. Then we put some activity in front of the sheet (e.g. we can use a point source at a large distance from the sheet), ensuring that all holes receive approximately the same amount of photons. With this set up, an image is acquired. Deviations between the image and the known dot pattern allow to compute $\Delta_x(x, y)$ and $\Delta_y(x, y)$.

4.4.2 Energy correction

Recall that the energy of the detected photon was computed as the sum of all PMT responses (section 4.1.3.1). Of course, the sum is dominated by the few PMTs close to the point of scintillation, the other PMT's receive very few or even no scintillation photons. Because the PMT-characteristics show small individual differences, the computed energy is position dependent. This results in small shifts of the computed energy spectrum with position. The deviations are systematic (they vary only very slowly in time), so they can be measured and stored, so that a position dependent correction can be applied:

$$E_{\text{corrected}}(x, y) = E(x, y) + \Delta_E(x, y) \tag{4.24}$$

It is important that the energy window is nicely symmetrical around the photopeak, small shifts of the window result in significant changes in the number of accepted photons (figure 4.19). Consequently, if no energy correction is applied, the sensitivity for acceptable photons would be position dependent.

The energy correction table needs to be rebuilt about every six months or after a major intervention. The situation is similar for single crystal and multicrystal systems.

4.4.3 Uniformity correction

When linearity and energy correction have been applied, the image of a uniform activity distribution should be a uniform image. In practice, there may still be small differences, due to small variations of the characteristics with position (e.g. crystal transparency, optical coupling between crystal and PMT etc). Most likely, those are simply differences in sensitivity,

so they do not produce deformation errors, only multiplicative errors. Again, these sensitivity changes can be measured and stored in order to correct them.

To measure the sensitivity we acquire an image of a uniform activity. For the gamma camera, one approach is to put a large homogeneous activity in front of the collimator. Another possibility is to remove the collimator and put a point source at a large distance in front of the collimator. Let us assume that the point source is exactly in front of the center of the detector, at a distance H from the crystal. The number of photons arriving in position (x, y) per second is then

$$\text{Number of counts per second} = \frac{A}{4\pi(H^2 + (x - x_0)^2 + (y - y_0)^2)} \tag{4.25}$$

where A is the strength of the source in Bq (Becquerel), and (x_0, y_0) is the center of the crystal. Thus, if we know the size of the detector and the uniformity error we will tolerate we can compute the required distance H. You can verify that $H = 5D$, if D is the length of the crystal diagonal and if we accept an error of 1%.

For the PET camera, the approach is similar. Either a uniform phantom is used as for the gamma camera, but then it must be (slowly) rotated in the gantry to acquire all projections in the same way. Or a non-uniform irradiation is applied (e.g. a small phantom in the center of the field of view), and the geometry is taken into account when computing the detector sensitivities.

The number of photons detected in every pixel is Poisson distributed. Recall from section 2.2 that the standard deviation of a Poisson number is proportional to the square root of the expected value. So if we accept an error (a standard deviation) of 0.5%, we need to continue the measurement until we have about 40000 counts per pixel. Computation of the uniformity correction matrix is straightforward:

$$\text{sensitivity\_corr}(x, y) = \frac{\text{mean}(I)}{I(x, y)} \tag{4.26}$$

where I is the acquired image.

Consequently, if a uniformity correction is applied to a planar image, one Poisson variable is divided by another one. The relative noise variance in the corrected image will be the sum of the relative noise variances on the two Poisson variables (see appendix 10.4 for estimating the error on a function of noisy variables).

4.4.4 Dead time correction

"Dead time" means that the system has a limited data processing capacity. It the data input is too high, a fraction of the data will be ignored because of lack of time. As a result, the measured count rate is lower than the true count rate, and the error increases with increasing true count rate. Figure 4.20 shows the measured count rate as a function of the true count rate for a dead time of 700 ns. The gamma camera and the PET camera both have a finite dead time, consisting of several components. To describe the dead time, we will assume that all contributions can be grouped in two components, a front-end component and a data-processing component.

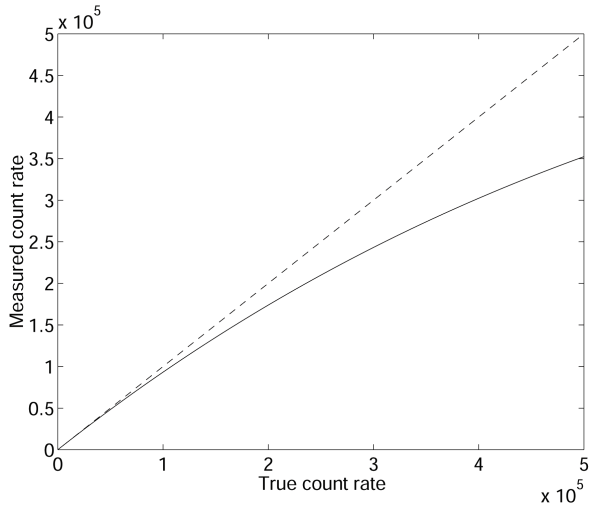

Figure 4.20: *Due to dead time, the measured count rate is lower than the true count rate for high count rates. The figure is for a dead time of 700 ns, the axes are in counts per second. Measured count rate is 20% low near 300000 counts per second.*

4.4.4.1 Front-end dead time

Recall (section 4.1.3.1) that the scintillation has a finite duration. E.g., the decay time of NaI(Tl) is 230 ns. The PMT response time adds a few ns. To have an accurate estimate of the energy, we need to integrate over a sufficient fraction of the PMT-outputs, such that most of the scintillation photons get the chance to contribute. Let us assume that we integrate over τ_1. Then the result is only correct if no other scintillation starts during that period. Otherwise, part of the second scintillation will contribute as well, the energy will be overestimated and the photon is rejected. Consequently, a photon is only detected if for a time τ_1 no other photon arrives. If we have a true count rate of R_0, then we expect $R_0\tau_1$ photons in the interval τ_1. Recall that the number of photons is Poisson distributed, so the probability of having 0 photons when $R_0\tau_1$ are expected is $\exp(-R_0\tau_1)$. Consequently, the count rate of accepted events will be

$$R_1 = R_0 e^{-R_0\tau_1}. \tag{4.27}$$

Notice that if R_0 is extremely large, R_1 goes to zero: the system is said to be *paralyzable*. Indeed, if the count rate is extremely high, every scintillation will be disturbed by the next one, and the camera will accept no events at all.

4.4.4.2 Data-processing dead time

When the event is accepted, the linearity correction must be applied and it must be stored somehow: the electronics either stores it in a list, and increments the pointer to the next available address, or it increments the appropriate location in the image. This takes time, let us say τ_2. If during that time another event occurs, it is simply ignored. Consequently, a fraction of the incoming events will be missed, because the electronics is not permanently available.

Assume that the electronics needs to store data at a rate of R_2 (unit s^{-1}). Each of these requires τ_2 s, so in every second, $R_2\tau_2$ are already occupied for processing data. The efficiency

of the system is therefor $1 - \tau_2 R_2$, which gives us the relation between incoming count rate R_1 and processed count rate R_2:

$$R_2 = (1 - R_2\tau_2)R_1 \tag{4.28}$$

Rearranging this to obtain R_2 as a function of R_1 results in

$$R_2 = \frac{R_1}{1 + R_1\tau_2}. \tag{4.29}$$

This function is monotonically increasing with upper limit $1/\tau_2$ (as you would expect: this is the number of intervals τ_2 in one second). Consequently, the electronics is non-paralyzable. You cannot paralyze it, but you can saturate it: if you put in too much data, it simply performs at maximum speed ignoring all the rest.

4.4.4.3 Effective dead time

Combining (4.27) and (4.29) tells us the acceptance rate for an incident photon rate of R_0:

$$R_0 = \frac{R_0 e^{-R_0\tau_1}}{1 + R_0 e^{-R_0\tau_1}\tau_2}. \tag{4.30}$$

If the count rates are small compared to the dead times (such that $R_x\tau_y$ is small), we can introduce an approximation to make the expression simpler. The approximation is based on the following relations, which are acceptable if x is small:

$$e^{-x} \simeq e^{-0} + x \left.\frac{de^{-x}}{dx}\right|_{x=0} = 1 - x \tag{4.31}$$

$$= \frac{1+x}{1+x} - x = \frac{1 + x - x - x^2}{1+x} \simeq \frac{1}{1+x}. \tag{4.32}$$

Applying this to (4.27) and (4.29) yields:

$$R_1 \simeq R_0(1 - R_0\tau_1) \tag{4.33}$$

$$R_2 \simeq R_1(1 - R_1\tau_2) \tag{4.34}$$

Combining both equation and deleting higher order terms results in

$$R_2 \simeq R_0\left(1 - R_0(\tau_1 + \tau_2)\right) \tag{4.35}$$

$$\simeq R_0 e^{-R_0(\tau_1+\tau_2)} \tag{4.36}$$

So if the count rate is relatively low, there is little difference between paralyzable and non-paralyzable dead time, and we can obtain a global dead time for the whole machine by simply adding the contributing dead times. Most clinical gamma cameras and all PET systems provide dead time correction based on some dead time model. However, for very high count rates the models are no longer valid and the correction will not be accurate.

4.4.5 Random coincidence correction

The random coincidence was defined in section 4.2.2: two non-related events may be detected simultaneously and be interpreted as a coincidence. There is no way to discriminate between true and random coincidences. However, there is a "simple" way to measure a reliable estimate of the number of randoms along every projection line. This is done via the delayed window technique.

The number of randoms (equation (4.21)) was obtained by squaring the probability to measure a single event during a time window τ. In the delayed window technique, an additional time window is used, which is identical in length to the normal one but delayed over a short time interval. The data for the second window are monitored with the same coincidence electronics and algorithms as used for the regular coincidence detection, ignoring the delay between the windows. If a coincidence is obtained between an event in the normal and an event in the delayed window, then we know it has to be a random coincidence: because of the delay the events cannot be due to the same annihilation. Consequently, with regular windowing we obtain *trues + randoms1*, and with the delayed method we obtain a value *randoms2*. The values *randoms1* and *randoms2* are not identical because of Poisson noise, but at least their expectations are identical. Subtraction will eliminate the bias, but will increase the variance on the estimated number of trues. See appendix 10.4 for some comments on error propagation.

Chapter 5

Image formation

5.1 Introduction

The tomographic systems are shown (once again) in figure 5.1. They have in common that they perform an indirect measurement of what we want to know: we want to obtain information about the distribution in every point, but we acquire information integrated over projection lines. So the desired information must be computed from the measured data. In order to do that, we need a mathematical function that describes (simulates) what happens during the acquisition: this function is an operator that computes measurements from distributions. If we have that function, we must derive its inverse: this will be an operator that computes distributions from measurements.

In the previous chapters we have studied the physics of the acquisition, so we are ready to compose the acquisition model. We can decide to introduce some approximation to obtain a simple acquisition model, hoping that deriving the inverse will not be too difficult. The disadvantage is that the inverse operator will not be an accurate inverse of the true acquisition process, so the computed distribution will not be identical to the true distribution. On the other hand, if we start from a very detailed acquisition model, inverting it may become mathematically and/or numerically intractable.

For a start, we will assume that collimation is perfect and that there is no noise and no

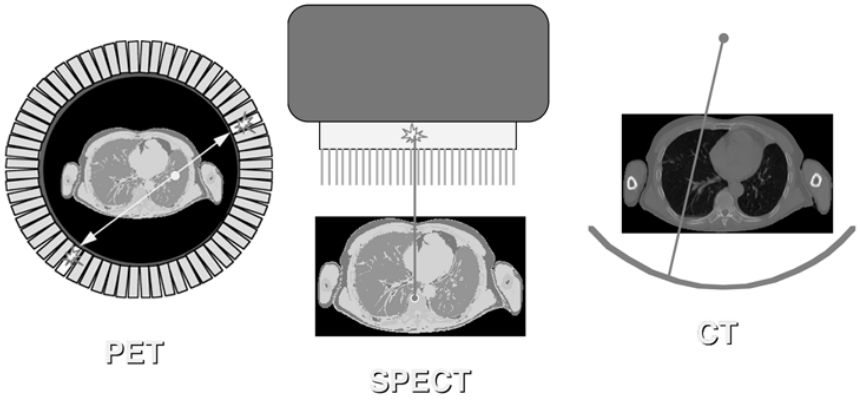

Figure 5.1: *Information acquired along lines in the PET camera, the gamma camera and the CT-scanner.*

scatter. We will only consider the presence of radioactivity and attenuating tissue.

Consider a single projection line in the CT-camera. We define the s-axis such that it coincides with the line. The source is at $s = a$, the detector at $s = b$. A known amount of photons t_0 is emitted in a towards the detector element at b. Only $t(b)$ photons arrive, the others have been eliminated by attenuation. If $\mu(s)$ is the linear attenuation coefficient in the body of the patient at position s, we have (see also eq (3.3))

$$t(b) = t_0 e^{-\int_a^b \mu(s)ds}. \tag{5.1}$$

The exponential gives the probability that a photon emitted in a towards b is not attenuated and arrives safely in b.

For a gamma camera measurement, there is no transmission source. Instead, the radioactivity is distributed in the body of the patient. We must integrate the activity along the s-axis, and attenuate every point with its own attenuation coefficient. Assuming again that the patient is somewhere between the points a and b on the s-axis, then the number of photons q arriving in b is:

$$q(b) = \int_a^b \lambda(s) e^{-\int_s^b \mu(\xi)d\xi} ds, \tag{5.2}$$

where $\lambda(s)$ is the activity in s. For the PET camera, the situation is similar, except that it must detect both photons. Both have a different probability of surviving attenuation. Since the detection is only valid if both arrive, and since their fate is independent, we must multiply the survival probabilities:

$$\begin{aligned} q(b) &= \int_a^b \lambda(s) e^{-\int_a^s \mu(\xi)d\xi} e^{-\int_s^b \mu(\xi)d\xi} ds & (5.3)\\ &= e^{-\int_a^b \mu(\xi)d\xi} \int_a^b \lambda(s) ds. & (5.4) \end{aligned}$$

In CT, only the attenuation coefficients are unknown. In emission tomography, both the attenuation and the source distribution are unknown. In PET, the attenuation is the same for the entire projection line, while in single photon emission, it is different for every point on the line.

It can be shown (and we will do so below) that it is possible to reconstruct the planar distribution, if all line integrals (or projections) through a planar distribution are available. As you see from figure 5.1, the gamma camera and the CT camera have to be rotated if all possible projections must be acquired. In contrast, a PET ring detector acquires all projections simultaneously (with "all" we mean here a sufficiently dense sampling).

5.2 Planar imaging

Before discussing how the activity and/or attenuation distributions can be reconstructed, it is interesting to have a look at the raw data. For a gamma camera equipped with parallel hole collimator, the raw data are naturally organized as interpretable images, "side views" of the tracer concentration in the transparent body of the patient. For a PET camera, the raw data can be organized in sets of parallel projections as well, as shown in figure 5.2.

Figure 5.3 shows a planar whole body image, which is obtained by slowly moving the gamma camera over the patient and combining the projections into a single image. In this

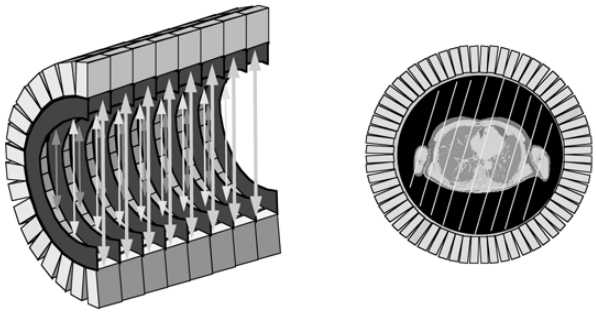

Figure 5.2: *Raw PET data can be organized in parallel projections, similar as in the gamma camera.*

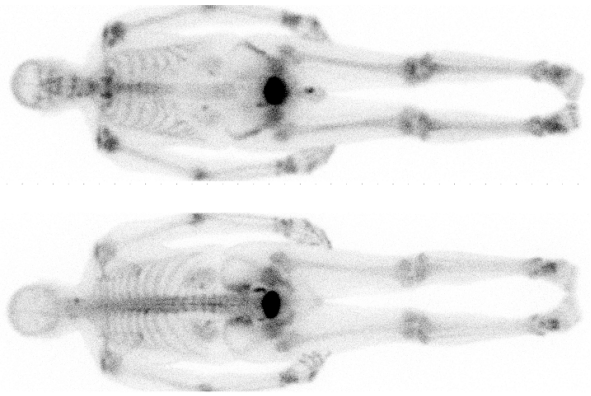

Figure 5.3: ^{99m}*Tc-MDP study acquired on a dual head gamma camera. Detector size is about 40 × 50 cm, the whole body images are acquired with slow translation of the patient bed. MDP accumulates in bone, allowing visualization of increased bone metabolism. Due to attenuation, the spine is better visualized on the posterior image.*

case, a gamma camera with two opposed detector heads was used, so the posterior and anterior views are acquired simultaneously. There is no rotation, these are simply raw data, but they provide valuable diagnostic information. A similar approach is applied in radiological studies: with the CT tomographic images can be produced, but the planar images (e.g. thorax X-ray) already provide useful information.

The study duration should not last too long because of patient comfort. A scanning time of 15 or even 30 min is reasonable, more is only acceptable if unavoidable. If that time is used to acquire a single or a few planar images, there will be many photons per pixel resulting in good signal to noise ratio, but there is no depth information. If the same time is used to acquire many projections for tomographic reconstruction, then there will be few counts per pixel and more noise, but we can reconstruct a three-dimensional image of the tracer concentration. The choice depends on the application.

In contrast, planar PET studies are very uncommon, because most PET-cameras are acquiring all projections simultaneously, so the three-dimensional image can always be reconstructed.

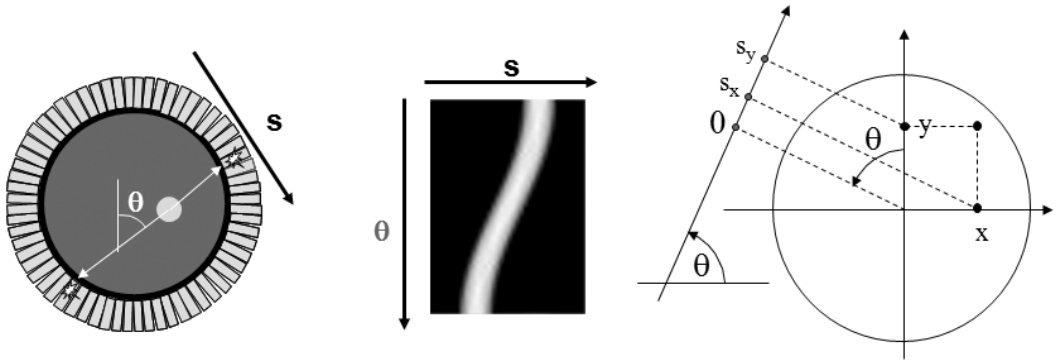

Figure 5.4: *A sinogram is obtained by storing the 1D parallel projections as the rows in a matrix (or image). Left: the camera with a radioactive disk. Center: the corresponding sinogram. Right: the conventions, θ is positive in clockwise direction.*

5.3 2D Tomography

In this section, we will study the reconstruction of a planar distribution from its line integrals. The data are two-dimensional: a projection line is completely specified by its angle and its distance to the center of the field of view. A two-dimensional image that uses d as the column coordinate and θ as the row coordinate is called a *sinogram*. Figure 5.4 shows that the name is well chosen: the sinogram of point source is zero everywhere except on a sinusoidal curve. The sinogram in the figure is for 180°. A sinogram for 360°shows a full period of the sinusoidal curve. It is easy to show that, using the conventions of figure 5.4, $d(x, y) = d_x + d_y = x \cos\theta + y \sin\theta$, so the non-zero projections in the sinogram are indeed following a sinusoidal curve.

5.3.1 2D filtered backprojection

Filtered backprojection (FBP) is the mathematical inverse of an idealized acquisition model: it computes a two-dimensional continuous distribution from ideal projections. In this case, "ideal" means the following:

- The distribution has a finite support: it is zero, except within a finite region. We assume this region is circular (we can always do that, since the distribution must not be non-zero within the support). We assume that this region coincides with the field of view of the camera. We select the center of the field of view as the origin of a two-dimensional coordinate system.
- Projections are continuous: the projection is known for every angle and for every distance from the origin. Projections are ideal: they are unweighted line integrals (so there is no attenuation and no scatter, the PSF is a Dirac impulse and there is no noise).
- The distribution is finite everywhere. That should not be a problem in real life.

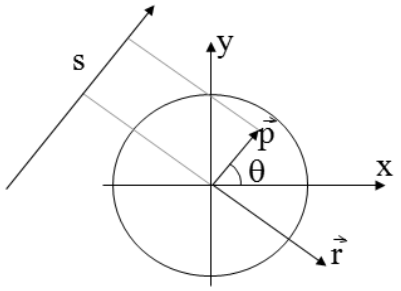

Figure 5.5: *The projection line as a function of a single parameter r. The line consists of the points $\vec{p}+\vec{r}$, with $\vec{p}=(d\cos\theta, d\sin\theta)$, $\vec{r}=(r\cos\theta, -r\sin\theta)$.*

5.3.1.1 Projection

Such an ideal projection q is then defined as follows (using the conventions of figure 5.4):

$$q(d,\theta) = \int_{(x,y)\in \text{ projection line}} \lambda(x,y)dxdy \tag{5.5}$$

$$= \int_{-\infty}^{\infty} \lambda(d\cos\theta + r\sin\theta, d\sin\theta - r\cos\theta)dr. \tag{5.6}$$

The point $(d\cos\theta, d\sin\theta)$ is the point on the projection line closest to the center of the field of view. By adding $(r\sin\theta, -r\cos\theta)$ we take a step r along the projection line (fig 5.5).

5.3.1.2 The Fourier theorem

The Fourier theorem states that there is a simple relation between the one-dimensional Fourier transform of the projection $q(d,\theta)$ and the two-dimensional Fourier transform of the distribution $\lambda(x,y)$.

The Fourier theorem is very easy to prove for projection along the y-axis (so $\theta = 0$). Because the y-axis may be chosen arbitrarily, it holds for any other projection angle (but the notation gets more elaborate). In the following Q is the 1-D Fourier transform (in the first coordinate) of q. Because the angle is fixed and zero, we have $q(d,\theta) = q(d,0) = q(x,0) = q(x)$.

$$q(x) = \int_{-\infty}^{\infty} \lambda(x,y)dy \tag{5.7}$$

$$Q(\nu_x) = \int_{-\infty}^{\infty} q(x)e^{-j2\pi\nu_x x} \tag{5.8}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \lambda(x,y)e^{-j2\pi\nu_x x}dxdy. \tag{5.9}$$

Let us now compute the 2-D Fourier transform Λ of λ:

$$\Lambda(\nu_x,\nu_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \lambda(x,y)e^{-j2\pi(\nu_x x+\nu_y y)}. \tag{5.10}$$

So it immediately follows that $\Lambda(\nu_x, 0) = Q(\nu_x)$. Formulated for any angle θ this becomes:

$$\Lambda(\nu\cos\theta, \nu\sin\theta) = Q(\nu,\theta). \tag{5.11}$$

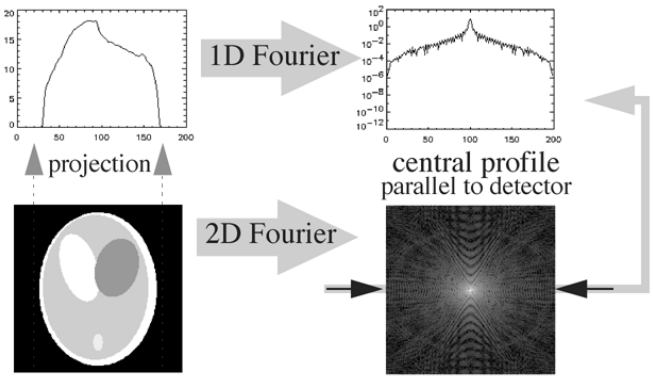

Figure 5.6: *The Fourier theorem.*

In words: the 1-D Fourier transform of the projections acquired for angle θ is identical to a central profile along the same angle through the 2-D Fourier transform of the original distribution (fig 5.6).

The Fourier theorem directly leads to a reconstruction algorithm by simply digitizing everything: compute the 1D FFT transform of the projections for every angle, fill the 2D FFT image by interpolation between the radial 1D FFT profiles, and compute the inverse 2D FFT to obtain the original distribution.

Obviously, this will only work if we have enough data. Every projection produces a central line through the origin of the 2D frequency space. We need to have an estimate of the entire frequency space, so we need central lines over at least 180 degrees (more is fine but redundant). Figure 5.7 shows clinical raw emission data acquired for tomography. They can be organized either as projections or as a sinograms. There are typically in the order of 100 (SPECT) or a few hundreds (PET) of projection angles. The figure shows 9 of them. During a clinical study, the gamma camera automatically rotates around the patient to acquire projections over 180°or 360°. Because the spatial resolution deteriorates rapidly with distance to the collimator, the gamma camera not only rotates, it also moves radially as close as possible to the patient to optimize the resolution.

As mentioned before, the PET camera consisting of detector rings measures all projection lines simultaneously, no rotation is required.

5.3.1.3 Backprojection

This Fourier-based reconstruction works, but usually an alternative expression is used, called "filtered backprojection". To explain it, we must first define the operation *backprojection*:

$$\begin{aligned} b(x,y) &= \int_0^\pi q(x\cos\theta + y\sin\theta, \theta)d\theta \\ &= \text{Backproj}\left(q(x,\theta)\right). \end{aligned} \tag{5.12}$$

During backprojection, every projection value is uniformly distributed along its projection line. Since there is exactly one projection line passing through (x, y) for every angle, the operation involves integration over all angles. This operation is *NOT* the inverse of the projection. As we will see later on, it is the *transpose*. Figure 5.8 shows the backprojection image using a logarithmic gray value scale. The backprojection image has no zero pixel values!

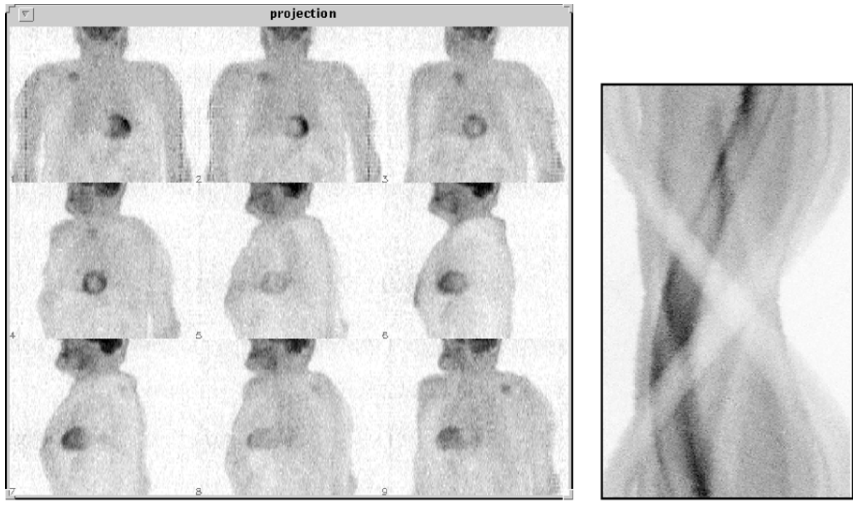

Figure 5.7: *Raw PET data, organized as projections (a) or as a sinogram (b). There are typically a few hundred projections, one for each projection angle, and several tens to hundred sinograms, one for each slice through the patient body*

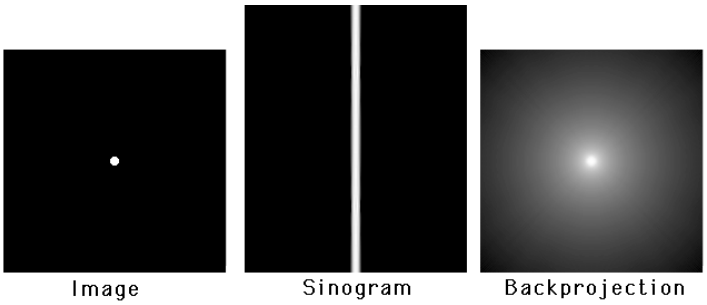

Figure 5.8: *A point source, its sinogram and the backprojection of the sinogram (logarithmic gray value scale).*

Let us compute the values of the backprojection image. Because the point is in the origin, the backprojection must be radially symmetrical. The sinogram is zero everywhere except for $d = 0$ (fig 5.8), where it is constant. So in the backprojection we only have to consider the central projection lines, the other lines contribute nothing to the backprojection image. Consider the line integral along a circle around the center (that is, the sum of all values on the circle). The circle intersects every central projection line twice, so the total activity the circle receives during backprojection is

$$\text{total circle value} = 2\int_0^{\pi} q(0,\theta)d\theta = 2q\pi. \tag{5.13}$$

So the line integral over a circle around the origin is a constant. The value in every point of a circle with radius R is $2q\pi/(2\pi R) = q/R$. Since projection followed by backprojection is a linear operation, we have actually computed to point spread function of that operation.

5.3.1.4 Filtered backprojection

Filtered backprojection follows directly from the Fourier theorem:

$$\begin{aligned}
\lambda(x,y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Lambda(\nu_x,\nu_y)e^{j2\pi(\nu_x x+\nu_y y)}d\nu_x d\nu_y && (5.14)\\
&\Downarrow \text{ Polar transform:} d\nu_x d\nu_y = |\nu| d\nu d\theta \\
&= \int_{-\infty}^{\infty} d\nu \int_0^{\pi} |\nu| d\theta \Lambda(\nu\cos\theta, \nu\sin\theta)e^{j2\pi\nu(x\cos\theta+y\sin\theta)} && (5.15)\\
&\Downarrow \text{ Fourier theorem and switching integral signs} \\
&= \int_0^{\pi} d\theta \int_{-\infty}^{\infty} |\nu| d\nu Q(\nu,\theta)e^{j2\pi\nu(x\cos\theta+y\sin\theta)} && (5.16)\\
&\Downarrow \text{ Apply definition of backprojection} \\
&= \text{Backproj}\left(\int_{-\infty}^{\infty}|\nu|Q(\nu,\theta)e^{j2\pi\nu x}\right) && (5.17)\\
&= \text{Backproj}\left(\text{Rampfilter}\left(q(x,\theta)\right)\right). && (5.18)
\end{aligned}$$

The ramp filter is defined as the sequence of 1D Fourier transform, multiplication with the "ramp" $|\nu|$ and inverse 1D Fourier transform. To turn it into a computer program that can be applied to a real measurement, everything is digitized and 1D Fourier is replaced with 1D FFT. Although exact mathematical treatment is rather subtle, the ramp filter can also be implemented in the Fourier domain:

$$\lambda(x,y) = \text{Backproj}\left(\text{inverse-Fourier-of-Rampfilter} * q(x,\theta)\right). \quad (5.19)$$

One can regard the ramp filter as a high pass filter which is designed to undo the blurring caused by backprojecting the projection, where the blurring mask is the point spread function computed in section 5.3.1.3.

Mathematically, filtered backprojection is identical to Fourier-based reconstruction, so the same conditions hold: e.g. we need projections over 180 degrees. Note that after digitization, the algorithms are no longer identical. The algorithms have different numerical properties and will produce slightly different reconstructions. Similarly, implementing the ramp filter in the Fourier domain or as a spatial convolution will produce slightly different results after digitization.

5.3.1.5 Attenuation correction

From a CT-measurement, we can directly compute a sinogram of line integrals. Starting from eq (5.1) we obtain:

$$\ln\frac{t_0}{t(b)} = \int_a^b \mu(s)ds. \quad (5.20)$$

CT measurements have low noise and narrow PSF, so filtered backprojection produces good results in this case. For emission tomography, the assumptions are not valid. Probably the most important deviation from the FBP-model is the presence of attenuation.

Attenuation cannot be ignored. E.g. at 140 keV, every 5 cm of tissue eliminates about 50% of the photons. So in order to apply filtered backprojection, we should first correct for the effect of attenuation. In order to do so, we have to know the effect. This knowledge can be obtained with a measurement. Alternatively, we can in some cases obtain a reasonable estimate of that effect from the emission data only.

In order to measure attenuation, we can use an external radioactive source rotating around the patient, and use the SPECT or PET system as a CT-camera to do a transmission measurement. If the external source emits N_0 photons along the x-axis, the expected fraction of photons making it through the patient is:

$$\frac{N(d)}{N_0} = e^{-\int_a^d \mu(x)dx}. \tag{5.21}$$

This is identical to the attenuation-factor in the PET-projection, equation (5.4). So we can correct for attenuation by multiplying the emission measurement $q(d)$ with the correction factor $N_0/N(d)$. For SPECT (equation (5.2)), this is not possible.

In many cases, we can assume that attenuation is approximately constant (with known attenuation coefficient μ) within the body contour. Often, we can obtain a fair body contour by segmenting a reconstruction obtained without attenuation correction. In that case, (5.21) can be computed from the estimated attenuation image.

Bellini has adapted filtered backprojection for constant SPECT-like attenuation within a known contour, so in this particular case, we can obtain attenuation corrected images. But there is no extension of filtered backprojection to deal with attenuation in general.

5.3.2 Iterative Reconstruction

Filtered backprojection is nice and fast, but not very flexible. The actual acquisition differs considerably from the ideal projection model, and this deviation causes reconstruction artifacts. It has been mentioned that attenuation correction in SPECT may be problematic, but there are other deviations from the ideal line integral: the measured projections are not continuous but digital, they contain a significant amount of noise (Poisson noise), the point spread function is not a Dirac impulse and Compton scatter may contribute significantly to the measurement.

That is why many researchers have been studying iterative algorithms. The nice thing about these algorithms is that they are not based on mathematical inversion. All they need is a mathematical model of the acquisition, not its inverse. Such a forward model is much easier to derive and program. That forward model allows the algorithm to *evaluate* any reconstruction, and to *improve* it based on that evaluation. If well designed, iterative application of the same algorithm should lead to continuous improvement, until the result is "good enough".

There are many iterative algorithms, but they are rather similar, so explaining one should be sufficient. In the following, the maximum-likelihood expectation-maximization (ML-EM) algorithm is discussed, because it is currently by far the most popular one.

The algorithm is based on a Bayesian description of the problem. In addition, it assumes that both the solution and the measurement are digital. This is correct in the sense that both the solution and the measurement are stored in a digital way. However, the true tracer distribution is continuous, so the underlying assumption is that this distribution can be well described with a digital representation.

5.3.2.1 Bayesian approach

Assume that somehow we have computed the reconstruction Λ from the measurement Q. The likelihood that both the measurement and the reconstruction are the true ones ($p(Q \text{ and } \Lambda)$) can be rewritten as:

$$p(\Lambda|Q)p(Q) = p(Q|\Lambda)p(\Lambda). \tag{5.22}$$

It follows that

$$p(\Lambda|Q) = \frac{p(Q|\Lambda)p(\Lambda)}{p(Q)}. \tag{5.23}$$

(This expression is Bayes' rule). The function $p(\Lambda)$ is called *the prior*. It is the likelihood of an image, without taking into account the data. It is only based on knowledge we already have prior to the measurement. E.g. the likelihood of a patient image, clearly showing that the patient has no lungs and four livers is zero. The function $p(Q|\Lambda)$ is simply called *the likelihood* and gives the probability to obtain measurement Q assuming that the true distribution is Λ. The function $p(\Lambda|Q)$ is called *the posterior*. The probability $p(Q)$ is a constant value, since the data Q have been measured and are fixed during the reconstruction.

Maximizing $p(\Lambda|Q)$ is called the *maximum-a-posteriori (MAP)* approach. It produces "the most probable" solution. Note that this doesn't have to be the true solution. We can only hope that this solution shares sufficient features with the true solution to be useful for our purposes.

Since it is not trivial to find good mathematical expressions for the prior probability $p(\Lambda)$, it is often assumed to be constant, i.e. it is assumed that a priori all possible solutions have the same probability to be correct. Maximizing $p(\Lambda|Q)$ then reduces to maximizing the likelihood $p(Q|\Lambda)$, which is easier to calculate. This is called the *maximum-likelihood (ML)* approach.

5.3.2.2 The likelihood function for emission tomography

We have to compute the likelihood $p(Q|\Lambda)$, assuming that the reconstruction image Λ is available and represents the true distribution. In other words, how likely is it to measure Q with a PET or SPECT camera, when the true tracer distribution is Λ?

We start by computing what we would expect to measure. We have already done that, it is the attenuated projection from equations (5.2) and (5.4). However, we want a digital version here:

$$r_i = \sum_{j=1,J} c_{ij}\lambda_j, \quad i = 1, I. \tag{5.24}$$

Here, $\lambda_j \in \Lambda$ is the regional activity present in the volume represented by pixel j (since we have a finite number of pixels, we can identify them with a single index). The value r_i is the number of photons measured in detector position i (i combines the digitized coordinates (d, θ)). The value c_{ij} represents the sensitivity of detector i for activity in j. If we have good collimation, c_{ij} is zero everywhere, except for the j that are intersected by projection line i, so the matrix C is very sparse. This notation is very general, and allows us e.g. to take into account the finite acceptance angle of the mechanical collimator (which will increase the fraction of non-zero c_{ij}). If we know the attenuation coefficients, we can include them in the c_{ij}, and so on. Consequently, this approach is valid for SPECT and PET.

We now have for every detector two values: the expected value r_i and the measured value q_i. Since we assume that the data are samples from a Poisson distribution, we can compute the likelihood of measuring q_i, if r_i photons were expected (see eq. (2.8)):

$$p(q_i|r_i) = \frac{e^{-r_i} r_i^{q_i}}{q_i!}. \tag{5.25}$$

The history of one photon (emission, trajectory, possible interaction with electrons, possible detection) is independent of that of the other photons, so the overall probability is the product

of the individual ones:

$$p(Q|\Lambda) = \prod_i \frac{e^{-r_i} r_i^{q_i}}{q_i!}. \tag{5.26}$$

Obviously, this is going to be a very small number: e.g. $p(q_i = 15|r_i = 15) = 0.1$ and smaller for any other r_i. For larger q_i, the maximum p-value is even smaller. In a measurement for a single slice, we have in the order of 10000 detector positions, so the maximum likelihood value may be in the order of 10^{-10000}, which is zero in practice. We are *sure* the solution will be wrong. However, we hope it will be close enough to the true solution to be useful.

Maximizing (5.26) is equivalent to maximizing its logarithm, since the logarithm is monotonically increasing. When maximizing Λ, factors not depending on λ_j can be ignored, so we will drop $q_i!$ from the equations. The resulting log-likelihood function is

$$\begin{aligned} L(Q|\Lambda) &= \sum_i q_i \ln(r_i) - r_i & (5.27)\\ &= \sum_i q_i \ln(\sum_j c_{ij}\lambda_j) - \sum_j c_{ij}\lambda_j. & (5.28) \end{aligned}$$

It turns out that the Hessian (the matrix of second derivatives) is negative definite if the matrix c_{ij} has maximum rank. In practice, this means that the likelihood function has a single maximum, provided that a sufficient amount of different detector positions i were used.

5.3.2.3 Maximum-Likelihood Expectation-Maximization

Since we can compute the first derivative (the gradient), many algorithms can be devised to maximize L. A straightforward one is to set the first derivative to zero and solve for λ.

$$\frac{\partial L}{\partial \lambda_j} = \sum_i c_{ij} \left(\frac{q_i}{\sum_j c_{ij}\lambda_j} - 1 \right) = 0, \forall i = 1, I. \tag{5.29}$$

This involves inversion of a matrix of $I \times J$ elements (in the order of 10^8), which is impractical.

Iterative optimization, such as a gradient ascent algorithm, is a suitable alternative. Starting with an arbitrary image Λ, the gradient for every λ_j is computed, and a value proportional to that gradient is added. Gradient ascent is robust but can be very slow, so more sophisticated algorithms have been investigated.

A very nice and simple algorithm with guaranteed convergence is the *expectation- maximization (EM)* algorithm. Although the resulting algorithm is simple, the underlying theory is not. In the following we simply state that convergence is proved and only show what the EM algorithm does. The interested reader is referred to appendix 10.6 for some notes on convergence.

Expected value of Poisson variables, given a single measurement

The iterative algorithm described below makes use of the expected value of a Poisson variable that contributes to a measurement. This section shows how that value is computed. Consider the experiment of fig 5.9: two vials containing a known amount of radioactivity are put in front of a detector. Assume that we know the efficiency of the detector and its sensitivity for the two vials, so that we can compute the expected amount of photons that each of the vials will contribute during a measurement. The expected count is $\bar{a}$ for vial 1

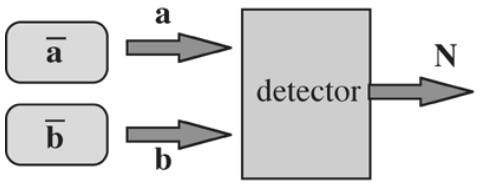

Figure 5.9: *A detector measures the counts emitted by two sources.*

and $\bar{b}$ for vial 2. Now a single experiment is carried out, and N counts are measured by the detector. Question: how many photons a and b were emitted by each of the vials?

A-priori, we would expect $\bar{a}$ photons from vial 1 and $\bar{b}$ photons from vial 2. But then the detector should have measured $\bar{a}+\bar{b}$ photons. In general, $N \neq \bar{a}+\bar{b}$ because of Poisson noise. So the measurement N supplies additional information, which we must use to improve our expectations about a and b. The answer is computed in appendix 10.5. The expected value of a, given N is:

$$E(a|a+b=N) = \bar{a}\frac{N}{\bar{a}+\bar{b}}, \tag{5.30}$$

and similar for b. So if more counts N were measured than the expected $\bar{a}+\bar{b}$, the expected value of the contributing sources is "corrected" with the same factor. Extension to more than two sources is straightforward.

The complete variables

First, we introduce a so-called set of "complete" variables $X = \{x_{ij}\}$, where x_{ij} is the (unknown) number of photons that have been emitted in j and detected in i. The x_{ij} are not observable, but if they are known, any observable variable can be computed. Obviously, the expected value of x_{ij}, given Λ is

$$E(x_{ij}|\Lambda) = c_{ij}\lambda_j \tag{5.31}$$

We can now derive a log-likelihood function for the complete variables X, in exactly the same way as we did for L (the x_{ij} obey Poisson statistics, just as q_i). This results in:

$$L_x(X,\Lambda) = \sum_i \sum_j (x_{ij}\ln(c_{ij}\lambda_j) - c_{ij}\lambda_j) \tag{5.32}$$

The EM algorithm prescribes a two stage procedure (and guarantees that doing so leads to the maximization of both L_x and L):

1. Compute the function $E(L_x(X,\Lambda)|Q,\Lambda^{old})$. It is impossible to compute $L_x(X,\Lambda)$, since we don't know the values of x_{ij}. However, we can calculate its *expected* value, using the current estimate Λ^{old}. This is called the *E-step*.

2. Calculate a new estimate of Λ that maximizes the function derived in the first step. This is the *M-step*.

The E-step

The E-step yields the following expressions:

$$E(L_x(X,\Lambda)|Q,\Lambda^{old}) \;=\; \sum_i \sum_j (n_{ij}\ln(c_{ij}\lambda_j) - c_{ij}\lambda_j) \tag{5.33}$$

$$n_{ij} = c_{ij}\lambda_j^{old}\frac{q_i}{\sum_k c_{ik}\lambda_k^{old}} \tag{5.34}$$

Equation (5.33) is identical to equation (5.32), except that the unknown x_{ij} values have been replaced by their expected values n_{ij}. We would expect that n_{ij} equals $c_{ij}\lambda_j^{old}$. However, we also know that the sum of all $c_{ij}\lambda_j^{old}$ equals the number of measured photons q_i. This situation is identical to the problem studied in section 5.3.2.3, and equation (5.34) is the straightforward extension of equation (5.30) for multiple sources.

The M-step

In the M-step, we maximize this expression with respect to λ_j, by setting the partial derivative to zero:

$$\frac{\partial}{\partial\lambda_j}E(L_x(X,\Lambda)|Q,\Lambda^{old}) = \sum_i\left(\frac{n_{ij}}{\lambda_j} - c_{ij}\right) = 0 \tag{5.35}$$

So we find:

$$\lambda_j = \frac{\sum_i n_{ij}}{\sum_i c_{ij}} \tag{5.36}$$

Substitution of equation (5.34) produces the ML-EM algorithm:

$$\lambda_j^{new} = \frac{\lambda_j^{old}}{\sum_i c_{ij}}\sum_i c_{ij}\frac{q_i}{\sum_j c_{ij}\lambda_j^{old}} \tag{5.37}$$

Discussion

This equation has a simple intuitive explanation:

1. The ratio $q_i/\sum_j c_{ij}\lambda_j^{old}$ compares the measurement to its the expected value, based on the current reconstruction. If the ratio is 1, the reconstruction must be correct. Otherwise, it has to be improved.

2. The ratio-sinogram is backprojected. The digital version of the backprojection operation is

$$\text{backprojection of } f \text{ equals: } \sum_j c_{ij}f_j. \tag{5.38}$$

   We have seen the continuous version of the backprojection in (5.12). The digital version clearly shows that backprojection is the transpose of projection: the operations are identical, except that backprojection sums over j, while projection sums over i.

3. Finally, the backprojected image is normalized and multiplied with the current reconstruction image. It is clear that if the measured and computed sinograms are identical, the entire operation has no effect. If the measured projection values are higher than the computed ones, the reconstruction values tend to get increased.

Comparing with (5.29) shows that the ML-algorithm is really a gradient ascent method. It can be rewritten as

$$\lambda_j^{\text{new}} = \lambda_j + \frac{\lambda_j}{\sum_j c_{ij}}\frac{\partial L}{\partial\lambda_j}. \tag{5.39}$$

So the gradient is weighted by the current reconstruction value, which is guaranteed to be positive (negative radioactivity is meaningless). To make it work, we can start with any non-zero positive initial image, and iterate until the result is "good enough". Fig. 5.10 shows the

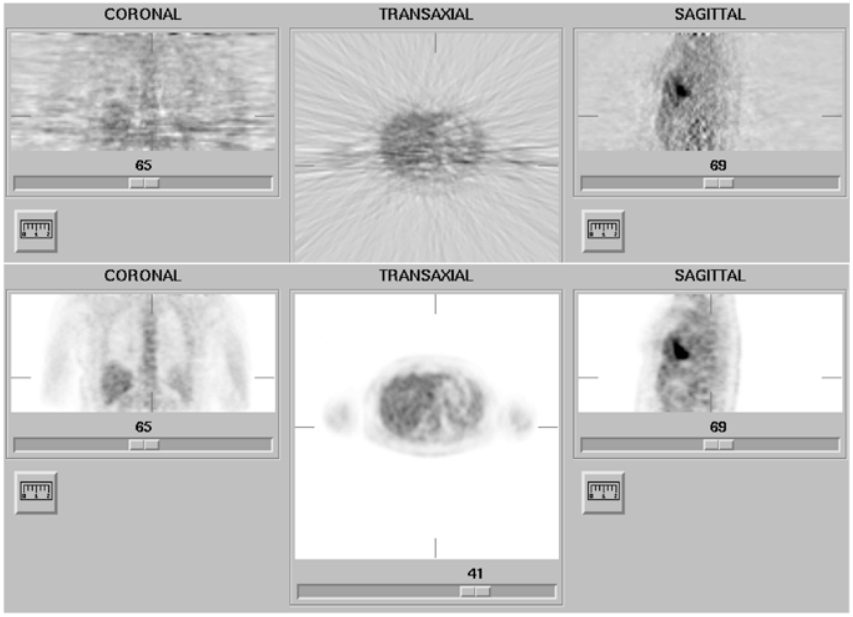

Figure 5.10: *Reconstruction obtained with filtered backprojection (top) and maximum-likelihood expectation-maximization (34 iterations) (bottom). The streak artifacts in the filtered backprojection image are due to the statistical noise on the measured projection (Poisson noise).*

filtered backprojection and ML-EM reconstructions from the same dataset. (The image shown is not the true maximum likelihood image, since iterations were stopped early as discussed below).

5.3.3 Regularization

Since the levels of radioactivity must be low, the number of detected photons is also low. As a result, the uncertainty due to Poisson noise is important: the data are often very noisy.

Note that Poisson noise is "white". This means that its (spatial) frequency spectrum is flat. Often this is hard to believe: one would guess it to be high frequency noise, because we see small points, no blobs or a shift in mean value. That is because our eyes are very good at picking up high frequencies: the blobs and the shift are there all right, we just don't see them.

Filtered backprojection has not been designed to deal with noise. The noise propagates unhampered into the reconstruction. In that process it is affected by the ramp filter, so the noise in the reconstruction is not white, it is truly high frequency noise. As a result, it is definitely *not* Poisson noise. Figure 5.10 clearly shows that Poisson noise leads to streak artifacts when filtered backprojection is used.

ML-EM does "know" about Poisson noise, but that does not allow it to separate the noise from the signal. In fact, MLEM attempts to compute how many of the detected photons have been emitted in each of the reconstruction pixels j. That must produce a noisy image, because photon emission is a Poisson process. What we really would like to know is the tracer concentration, which is not noisy.

Because our brains are not very good at suppressing noise, we need to do it with the

computer. Many techniques exist. One can apply simple smoothing, preferably in three dimensions, such that resolution is approximately isotropic. It can be shown that for every isotropic 3D smoothing filter applied after filtered backprojection, there is a 2D smoothing filter, to be applied to the measured data before reconstruction, which has the same effect. Since 2D filtering is faster than 3D filtering, it is common practice to smooth before filtered backprojection.

The same is not true for ML-EM: one should not smooth the measured projections, since that would destroy the Poisson nature. Instead, the images are often smoothed afterwards. Another approach is to stop the iterations before convergence. ML-EM has the remarkable feature that low frequencies converge faster than high ones, so stopping early has an affect similar to low-pass filtering.

Finally, one can go back to the basics, and in particular to the Bayesian expression (5.23). Instead of ignoring the prior, one can try and define some prior probability function that encourages smooth solutions. This leads to maximum-a-posteriori (MAP) algorithm. The algorithms can be very powerful, but they also have a lot of parameters and tuning them is a delicate task.

Although regularized (smoothed) images look much nicer than the original ones, they do not contain more information. In fact, they contain less information, since low-pass filtering kills high-frequency information and adds nothing instead! Deleting high spatial frequencies results in poorer resolution (wider point spread function), so excessive smoothing is ill advised if you hope to see small structures in the image.

5.3.4 Convergence

Filtered backprojection is a linear algorithm with shift invariant point spread function. That means that the effect of projection followed by filtered backprojection can be described with a single PSF, and that the reconstruction has a predictable effect on image resolution. Similarly, smoothing is linear and shift invariant, so the combined effect of filtered backprojection with low pass filtering has a known effect on the resolution.

In contrast to FBP, MLEM algorithm is non-linear and not shift invariant. Stated otherwise, if an object is added to the reconstruction, then the effect of projection followed by reconstruction is different for *all* objects in the image. In addition, two identical objects can be reconstructed differently because they are at a different position. In this situation, it is impossible to define a PSF, projection followed by MLEM-reconstruction cannot be modeled as a convolution. There is no way to predict the effect of MLEM on image resolution.

It is (in theory) possible to predict the resolution of the true maximum likelihood solution, but you need an infinite number of iterations to get there. Consequently, if iterations are stopped early, convergence may be incomplete. There is no way to tell that from the image, unless you knew a-priori what the image was: MLEM images always look nice. So stopping iterations early in patient studies has unpredictable consequences.

This is illustrated in a simulation experiment shown in figure 5.11. There are two point sources, one point source is surrounded by large active objects, the other one is in a cold region. After MLEM 20 iterations, the “free” point source has nearly reached its final value, while the other one is just beginning to appear. Even after 100 iterations there is still a significant difference. The difference in convergence affects not only the maximum count, but also the total count in a region around the point source.

Consequently, it is important to apply “enough” iterations, to get “sufficiently” close to

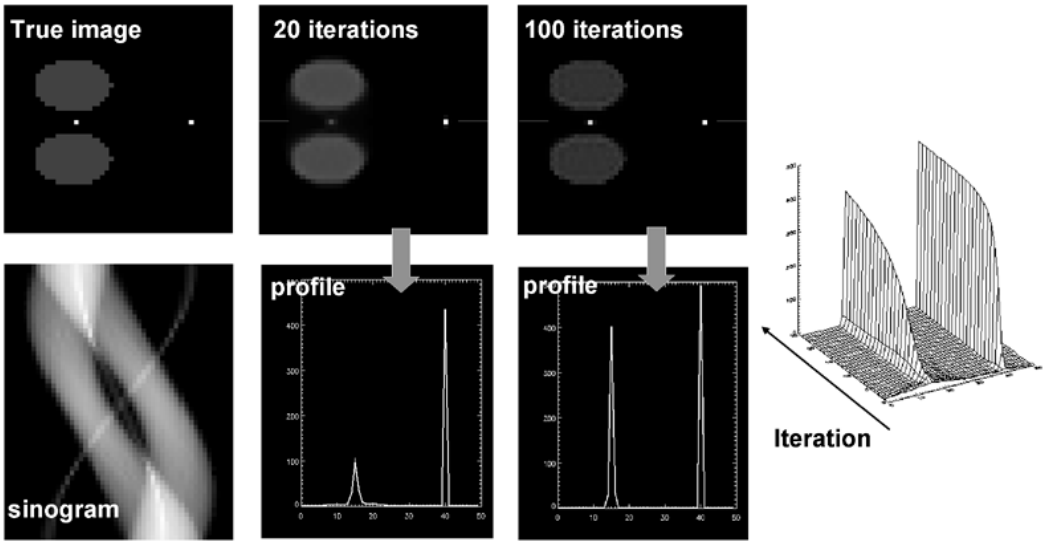

Figure 5.11: *Simulation designed to challenge convergence of MLEM: the point between the hot regions convergence very slowly relative to the other point. The maximum converges slowlier than the width, so counts are not preserved.*

the true ML solution. In the previous section it was mentioned that stopping iterations early has a noise-suppressing effect. This section shows that stopping too early is ill advised. For very noisy data, it is better to continue iterating and remove the noise afterwards with post-smoothing. In practice, sufficiently long means that at least 20 or 30 iterations are needed for typical SPECT and PET images. The number depends on the image size, so if sampling improves, the number of iterations must go up. If computer time is not a problem, it is recommended to apply more iterations (several hundreds) and regularize with a prior or with post-smoothing.

5.4 Fully 3D Tomography

Until now, we have only discussed the reconstruction of a single 2D slice from a set of 1D projection data. An entire volume can be obtained by reconstructing neighboring slices. This is what happens in SPECT with parallel hole collimation and in 2D PET.

In some configurations this approach is not possible and the problem has to be treated as fully three-dimensional one. There are many different geometries of mechanical collimators, and some of these acquire along lines that cannot be grouped in parallel sets. An example is the cone beam collimator (fig 4.5). Another example is 3D PET (section 4.2.2.4). Three approaches to fully 3D reconstruction are discussed below.

5.4.0.1 Filtered backprojection

Filtered backprojection can be extended to the fully 3D case. The data are backprojected along their detection lines, and then filtered to undo the blurring effect of backprojection. This is only possible if the sequence of projection and backprojection results in a shift-invariant point spread function. And that is only true if every point in the reconstruction volume is

intersected by the same configuration of measured projection lines.

This is often not the case in practice. Points near the edge of the field of view are intersected by fewer measured projection lines. In this case, the data may be completed by computing the missing projections as follows. First a subset of projections that meets the requirement is selected and reconstructed to compute a first (relatively noisy) reconstruction image. Next, this reconstruction is forward projected along the missing projection lines, to compute the missing data. Then, the computed and measured data are combined into a single set of data, that now meets the requirement of shift-invariance. Finally, this completed data set is reconstructed with 3D filtered backprojection.

5.4.0.2 ML-EM reconstruction

ML-EM can be directly applied to the 3D data set: the formulation is very general, the coefficients c_{ij} in (5.37) can be directly used to describe fully 3D projection lines.

In every iteration, we have to compute a projection and a backprojection along every individual projection line. As a result, the computational burden may become pretty heavy for a fully 3D configuration.

5.4.0.3 Fourier rebinning

Recently, an exact rebinning algorithm, called *Fourier rebinning*, was derived. It converts a set of 3D data into a set of 2D projections. These projections can be reconstructed with standard 2D filtered backprojection. It was also shown that the Poisson nature of the data is more or less preserved. Consequently, the resulting 2D set can be reconstructed successfully with the 2D ML-EM algorithm as well. (Fourier rebinning is based on a wonderful feature of the Fourier transform of the sinograms, hence its name.)

In practice, the exact rebinning algorithm is not used. Instead, one only applies an approximate expression derived from it, because it is much faster and sufficiently accurate for most configurations.

Chapter 6

The transmission scan

6.1 System design

To accurately reconstruct images from emission sinograms, information about the attenuation is needed. Recall that more information is needed for SPECT than for PET (section 5.3.1.5). For PET, the sinograms can be precorrected for attenuation and then be reconstructed with FBP. For SPECT, there is no general attenuation precorrection.

If reconstruction is done with MLEM, precorrection is not recommended since it destroys the Poisson nature of the data. It is better to include the attenuation factors in the coefficients c_{ij} in equation (5.37). Remember that MLEM deals with attenuation by simulating it, not by inverting it: there is no explicit attenuation *correction* in MLEM. The MLEM-algorithm simulates the effect of attenuation during projection, and will iterate until the computed projections are (nearly) equal to the measured ones. When attenuation is included the coefficients c_{ij} can be rewritten as

$$c_{ij} = w_{ij}a_{ij} \qquad \text{(SPECT)} \tag{6.1}$$

$$c_{ij} = w_{ij}a_i \qquad \text{(PET)}, \tag{6.2}$$

where w_{ij} is the detection sensitivity in absence of attenuation.

Several configurations have been devised to enable transmission scanning on the gamma camera. As a transmission source, point sources, line sources and sheet sources are being used. Figure 6.1 shows a scanning line source configuration. The line source is collimated both axially and transaxially. Collimation avoids photon emission along lines that are not accepted by the collimator in front of the crystal. Elimination of those photons reduces exposure to patient and personnel, and reduces the contribution of scattered photons.

The transmission isotope is selected to emit photons at an energy different from the emission tracer. Thus, emission and transmission photons can be separated, enabling simultaneous acquisition of transmission and emission projections. In the scanning line source configuration, an electronic window is moved in synchronization with the source. Photons outside the electronic window cannot have originated in the source, and must be due to scatter from emission photons (if the energy of the emission isotope is higher). Consequently, the electronic window improves the separation already achieved with the energy windows.

Figure 6.2 shows a typical PET transmission configuration. One or a few rotating rod sources are used. In theory, a single photon emitter could be used, since the position of the rod sources is known at all times. However, current systems usually use a positron emitter,

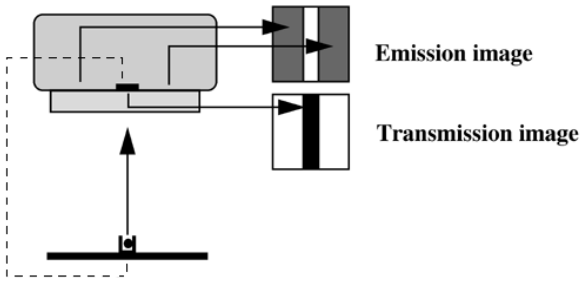

Figure 6.1: *Scanning transmission source in a gamma camera. An electronic window is synchronized with the source, improving separation of transmission and emission counts.*

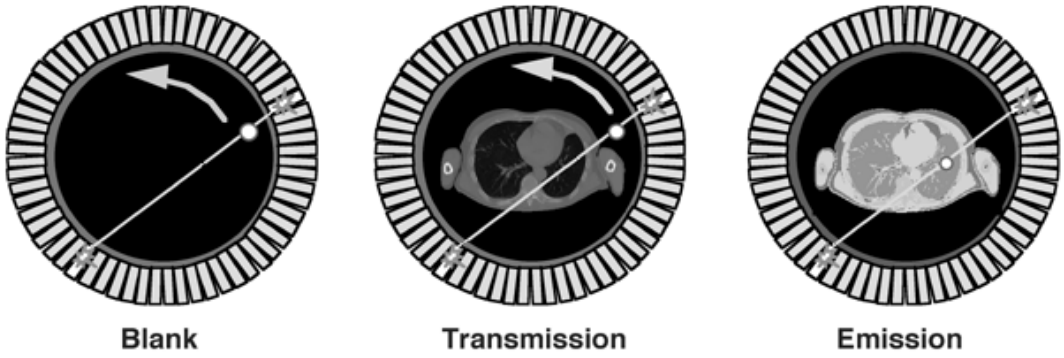

Figure 6.2: *Rotating transmission source in PET. As a reference, a blank scan is acquired daily.*

so separation based on energy windowing is not possible. Again, small electronic windows are used, which reduce the emission contamination with a factor of about 40. The remaining contamination can be estimated from an emission sinogram acquired immediately before or after the transmission scan.

For every projection line, we also need to know the number of photons emitted by the source. These values are measured during a blank scan. The blank scan is identical to the transmission scan, except that there is nothing in the field of view. Blank scans can be acquired over night, so they do not increase the study duration.

6.2 Attenuation correction

For PET, only the ratio between blank and transmission is required. However, it is useful to reconstruct the transmission image for two reasons. First, although the image quality is very poor compared to CT, it provides some anatomical reference which may be valuable to the physician. Second, the attenuation along a projection line may be computed from the reconstructed image. This improves the signal to noise ratio. Indeed, when computed from the reconstruction, the entire blank and transmission sinogram contribute to the estimated attenuation coefficient. In contrast, an estimate computed from the ratio of blank and transmission scans is based on a single blank and transmission pixel value.

For SPECT, the transmission measurement must be reconstructed, because the recon-

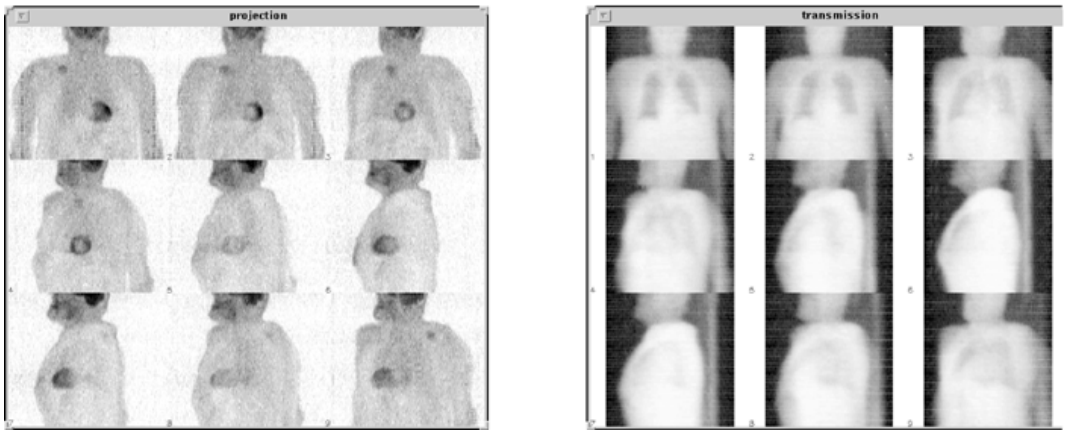

Figure 6.3: *Left: emission PET projections. Right: transmission projections of the same patient.*

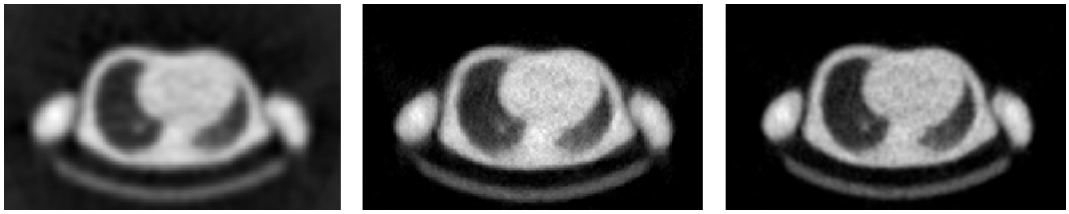

Figure 6.4: *Reconstruction of a PET transmission scan of* **10 min**. *Left: filtered backprojection; Center: MLTR; Right: MAP (Maximum a posteriori reconstruction).*

struction values are required to compute the coefficient c_{ij} in (5.37) or in other iterative algorithms.

All what has been said about reconstruction of emission scans can be done for transmission scans as well. However, there is an important difference. In emission tomography, the raw data are a weighted sum of the unknown reconstruction values. In transmission tomography, the raw data are proportional to the exponent of such a weighted sum. As a result of this difference, the MLEM algorithm cannot be applied directly, so several new ones have been presented in recent literature. Similarly as in the emission case, one can use a non-trivial prior distribution for the reconstruction images. In that case, the reconstruction is called a maximum-a-posteriori algorithm or MAP-algorithm (see section 5.3.2.1).

Figure 6.4 shows the reconstructions of a 10 min transmission scan. The same sinogram has been reconstructed with filtered backprojection, ML and MAP. Because the scan time is fairly long, image quality is reasonable for all algorithms.

Figure 6.5 shows the reconstructions of a 1 min transmission scan obtained with the same three algorithms. In this short scan, the noise is prominent. As a result, streak artifacts show up in the filtered backprojection image. The ML-image produces non-correlated noise with high amplitude. As argued in section 5.3.3 this can be expected, since the true number of photons attenuated during the experiment in every pixel is subject to statistical noise. And if that number is small, the relative noise amplitude is large. The MAP-reconstruction is much smoother, because the prior assigns a low probability to noisy solutions. This image is a compromise between what we know from the data and what we (believe to) know a priori.

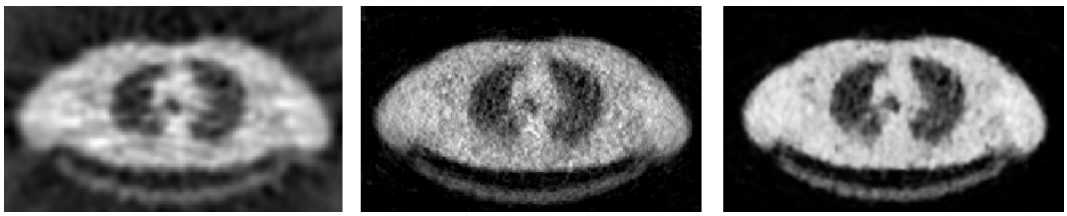

Figure 6.5: *Reconstruction of a PET transmission scan of* **1 min**. *Left: filtered backprojection; Center: MLTR; Right: MAP (Maximum a posteriori reconstruction).*

Chapter 7

Quality control

The photomultiplier tubes are analog devices, with characteristics that are influenced by the age and the history of the device. To produce correct images, some corrections are applied (section 4.4). The most important are the linearity correction and the energy correction. In most cameras, there is an additional correction, called the uniformity correction, taking care of remaining second order deviations. It is important to keep on eye on the camera performance and tune it every now and then. Failure to do so will lead to gradual deterioration of image quality.

When a camera is first installed, the client tests it extensively to verify that the system meets all required specifications (the acceptance test). The National Electrical Manufacturers Association (NEMA, "national" only holds for the USA) has defined standard protocols to measure the specifications. The standards are widely accepted, which allows to compare specifications from different vendors, and avoids discussions between customers and companies about acceptance test procedures. You can find information on *http://www.nema.org*.

This chapter gives an overview of tests which provide valuable information on camera performance. Many of these tests can be done quantitatively: they provide a number. *It is very useful to store the numbers and plot them as a function of time:* this helps detecting problems early, since gradual deterioration of performance is detected on the curve even before the performance deteriorates beyond the specifications.

Quality control testing is often tedious, and it is not productive on the short term: it costs time, and if you find an error, it also costs money. Most of the time, the camera is working well, so if you assume that the camera is working fine, you are probably right and you save time. This is why a quality control program needs continuous monitoring: if you don't insist on quality control measurements, the QC-program will silently die, and image quality will slowly deteriorate.

7.1 Gamma camera

For gamma camera quality control and acceptance testing, the *Nederlandse Vereniging voor Nucleaire Geneeskunde* has published a very useful book [3]. It provides detailed recipes for applying the measurement and processing procedures, but does not attempt to explain it, the reader is supposed to be familiar with nuclear medicine.

Figure 7.1 shows the influence of the corrections on the image of a uniform phantom. Without any correction, the photomultipliers are clearly visible as spots of increased intensity.

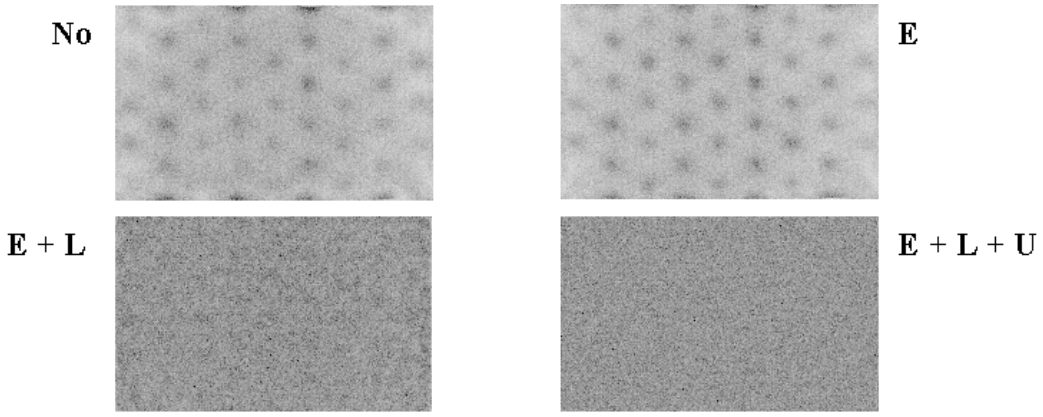

Figure 7.1: *Image of a uniform phantom. E = energy correction, L = linearity correction, U = uniformity correction.*

After energy correction this is still the case, but the response of the PMTs is more uniform. When in addition linearity correction is applied, the image is uniform, except for Poisson noise. Uniformity correction produces a marginal improvement which is hardly visible.

For most properties, specifications are supplied for the UFOV (usable field of view) and the CFOV (central field of view). The reason is that performance near the edges is always slightly inferior to performance in the center. Near the edges, the scintillation point is not nicely symmetrically surrounded by (nearly) identical PMTs. As a result, deviations are larger and much stronger corrections are needed, unavoidably leading to some degradation. The width of the CFOV is 75% of the UFOV width, both in x and y direction.

Quality control typically involves

- daily visual verification of image uniformity,
- weekly quantitative uniformity testing,
- monthly spatial resolution testing
- extensive testing every year or after major interventions.

The scheme can be adapted according to the strengths and weaknesses of the systems.

7.1.1 Planar imaging

7.1.1.1 Uniformity

Uniformity is evaluated by acquiring a uniform image. As a phantom, either a point source at large distance is measured without collimator, or a uniform sheet source is put on the collimator. It is recommended to do a quick uniformity test in the morning, to make sure that the camera seems to work fine before you start imaging the first patient. Figure 7.2 shows a sheet source image on a camera with a defect photomultiplier. Of course, you do not want to discover such a defect with a patient image.

For quantitative analysis, the influence of Poisson noise must be minimized by acquiring a large amount of counts (typically 10000) per pixel. Acquisition time will be in the order of

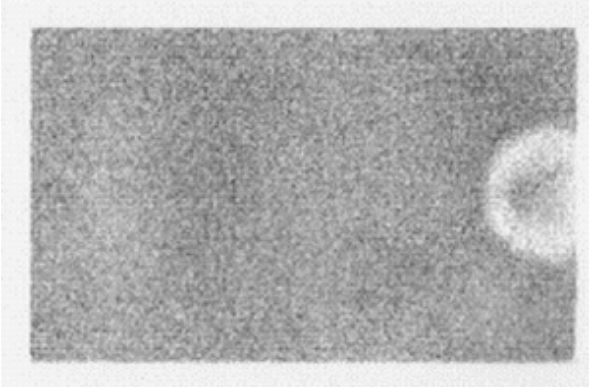

Figure 7.2: *Uniform image acquired on a gamma camera with a dead photomultiplier.*

an hour. Two parameters are computed from an image reduced to 64×64 pixels:

$$\text{Integral uniformity} = \frac{max - min}{max + min} \times 100\% \tag{7.1}$$

$$\text{Differential uniformity} = \max_{i=1..N} (\text{regional uniformity } i,) \tag{7.2}$$

where the regional uniformity is computed by applying (7.1) to a small line interval containing only 5 pixels, and this for all possible vertical and horizontal line intervals. The differential uniformity is always a bit smaller than the integral uniformity, and insensitive to gradual changes in image intensity.

In an image of 64×64 with 10000 counts per pixel, the integral uniformity due to Poisson noise only is about 4%, typical specifications are a bit larger, e.g. 4.5%.

To acquire a uniformity correction matrix, typically 45000 counts per pixel are acquired. *It is important that the camera is in good shape when a uniformity correction matrix is produced.* Figure 7.3 shows a flood source image acquired on a camera with a linearity correction problem. The linearity problem causes deformations in the image: counts are mispositioned. In a uniform image this leads to non-uniformities, in a line source image to deformations of the line. If we now acquire a uniformity correction matrix, the corrected flood image will of course be uniform. However, the line source image will still be deformed, and in addition the intensities of the deformed line will become worse if we apply the correction matrix! If uniformity is poor after linearity and energy correction, do not fix it with uniformity correction. Instead, try to figure out what happens or call the service people from the company.

The uniformity test is sensitive to most of the things that can go wrong with a camera, but not all. One undetected problem is deterioration of the pixel size, which stretches or shrinks the image in x or y direction.

7.1.1.2 Pixel size

The x and y coordinates are computed according to equation (4.1), so they are affected by the PMT characteristics, the amplification of the analogue signals prior to A/D conversion and possibly by the energy of the incoming photon, since that affects the PMT outputs.

Measuring the pixel size is simple: put two point sources at a known distance, acquire an image and measure the distance in the image. Sub-pixel accuracy is obtained by computing the mass center of the point source response. The precision will be better for larger distances.

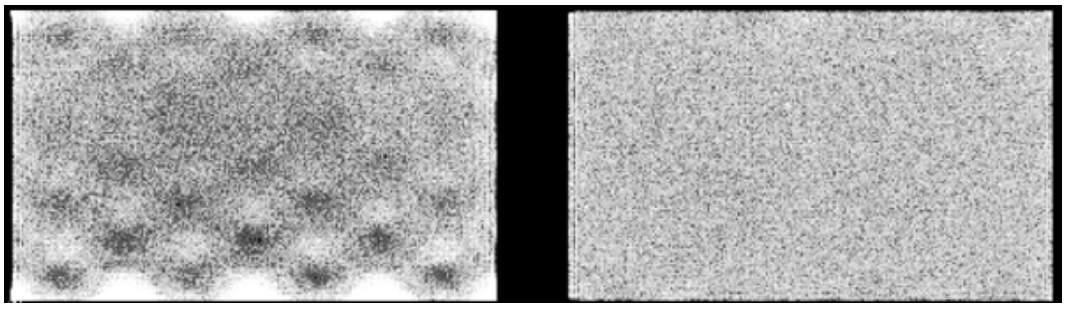

Figure 7.3: *Flood source (= uniform sheet source) image acquired on a dual head gamma camera, with a linearity correction problem in head 1 (left)*

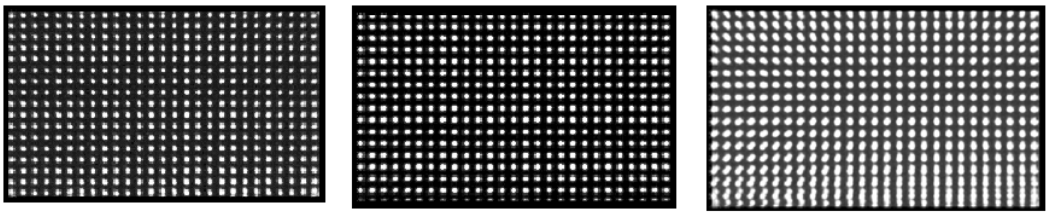

Figure 7.4: *Images of a dot phantom acquired on a dual head camera with a gain problem. Left: image acquired on head 1. Center: image simultaneously acquired on head 2. Right: superimposed images: head1 + mirror image of head2. The superimposed image is blurred because the heads have a different pixel size.*

The pixel size must be measured in x *and* in y direction, since usually each direction has its own amplifiers.

Figure 7.4 shows an example where the y-amplifier gain is wrong for one of the heads of a dual head camera (the x-gain is better but not perfect either). The error is found from this dot phantom measurement by superimposing both images. Since the images are recorded simultaneously, the dots from one head should fit those from the other. Because of the invalid y-pixel size in one of the heads there is position dependent axial blurring, which is worse for larger y values.

It is useful to verify if the pixel size is independent of the energy. This can be done with a ^{67}Ga point source, which emits photons of 93, 184 and 296 keV. Using three appropriate energy windows, three point source images are obtained. The points should coincide when the images are superimposed. Repeat the measurement at a few different positions (or use multiple point sources).

7.1.1.3 Spatial resolution

In theory, one could measure the FWHM of the PSF directly from a point source measurement. However, a point source affects only a few pixels, so the FWHM cannot be derived with good accuracy. Alternatively, one can derive it from line source measurements. The line spread function is the integral of the point spread function, since a line consists of many points on a row:

$$\mathrm{LSF}(x) = \int_{-\infty}^{\infty} \mathrm{PSF}(x, y) dy. \tag{7.3}$$

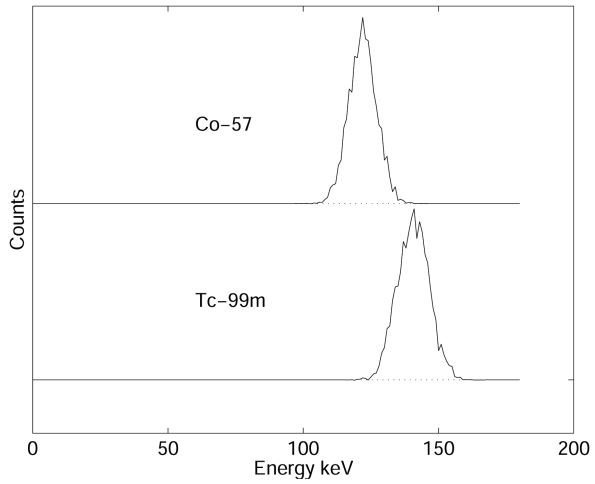

Figure 7.5: *(Simulated) energy spectra of Cobalt (^{57}Co, 122 keV) and technetium (^{99m}Tc, 140 keV).*

Usually, the PSF can be well approximated as a Gaussian curve. The LSF is then easy to compute:

$$\mathrm{LSF}(x) = \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} dy \tag{7.4}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}. \tag{7.5}$$

In these equations, we have assumed that the line coincides with the y-axis. So it is reasonable to assume that the FWHM of the LSF is the same as that of the PSF. On a single LSF, several independent FWHM measurements can be done and averaged to improve accuracy. Alternatively, it the line is positioned nicely in the x or y direction, one can first compute an average one-dimensional profile by summing image columns or rows and use that for LSF computations. Again, measurements for x and y must be made because the resolution can become anisotropic.

7.1.1.4 Energy resolution

Most gamma cameras can display the measured energy spectrum. Because it is not always clear where the origin of the energy axis is located, it is save to use two sources with a different energy, and calibrate the axis from the measurement. E.g. figure 7.5 shows two spectra, one for ^{99m}Tc (peak at 140 keV) and the other one for ^{57}Co (peak at 122 keV). The distance between the two peaks is 18 keV, and can directly be compared to the FWHM of the two spectra.

The FWHM is usually specified relative to the peak energy. So a FWHM of 10% for ^{99m}Tc means 14 keV.

7.1.1.5 Linearity

The linearity is measured with a line source phantom as shown in figure 7.6. This can be obtained in several ways. One is to use a lead sheet with long slids of 1 mm width, positioned

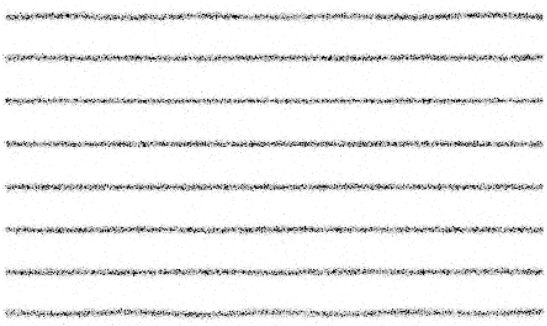

Figure 7.6: *Image of a line phantom acquired on a gamma camera with poor linearity correction.*

at regular distances (e.g. 3 cm distance between the slids). Remove the collimator (to avoid interference between the slids and the collimator septa, possibly resulting in beautiful but useless Moiré patterns), put the lead sheet on the camera (very carefully, the fragile crystal is easily damaged!) and irradiate with a uniform phantom or point source at large distance.

NEMA states that *integral linearity* is defined as the maximum distance between the measured line and the true line. The true line is obtained by fitting the known true lines configuration to the image.

The NEMA definition for the *differential linearity* may seem a bit strange. The procedure prescribes to compute several image profiles perpendicular to the lines. From each profile, the distances between consecutive lines is measured, and the standard deviation on these distances is computed. The maximum standard deviation is the *differential linearity*. This procedure is simple, but not 100% reproducible.

With current computers, it is not difficult to implement a better and more reproducible procedure. However, a good standard must not only produce a useful value, it must also be simple. If not, people may be reluctant to accept it, and if they do, they might make programming errors when implementing it. At the time the NEMA standards were defined, the procedure described above was a good compromise.

7.1.1.6 Dead time

The dead time measurements are based on some dead time model, for example equations (4.35) or (4.36). The effective dead time τ is the parameter we want to obtain. Usually, the exact amount of radioactivity is unknown as well, so there are two unknown variables. Consequently, we need at least two measurements to determine them. Many procedures can be devised. A straightforward one is to use a strong source with short half life, and acquire images while the activity decays. At low count rates the gamma camera is known to work well, so we assume that this part of the curve is correct (slope 1). Thus, we can compute what the camera should have measured, the true count rate, allowing us to draw the curve of figure 4.20. Then, τ can be computed, e.g. by fitting the model to the rest of the curve.

A faster method suggested in [3] is to prepare two sources with the same amount of radioactivity (difference less than 10%). Select the sources such that when combined the count rate is probably high enough to produce a noticeable dead time effect (otherwise the

subsequent analysis will be very sensitive to noise). Put one source on the camera and measure the count rate R_1. Put the second source on the camera and measure R_{12}. Remove the first source and measure R_2. This produces the following equations:

$$R_1 = R_1^* e^{-R_1^* \tau} \quad (7.6)$$
$$R_2 = R_2^* e^{-R_2^* \tau} \quad (7.7)$$
$$R_{12} = (R_1^* + R_2^*) e^{-(R_1^* + R_2^*)\tau}, \quad (7.8)$$

where the (unknown) true count rates are marked with an asterisk. If we did a good job preparing the sources, then we have $R_1 \simeq R_2$, so we can simplify the equations into

$$R = R^* e^{-R^* \tau} \quad (7.9)$$
$$R_{12} = 2R^* e^{-2R^* \tau}, \quad (7.10)$$

where we define $R = (R_1 + R_2)/2$. A bit of work (left as an exercise to the reader) leads to

$$\tau = \frac{2R_{12}}{(R_1 + R_2)^2} \ln \frac{R_1 + R_2}{R_{12}}. \quad (7.11)$$

Knowing τ, we can predict the dead time for any count rate, except for very high ones (which should be avoided at all times).

7.1.1.7 Sensitivity

Sensitivity is measured by recording the count rate for a known radioactive source. Because the measurement should not be affected by attenuation, NEMA prescribes to use a recipient with a large (10 cm) flat bottom, with a bit of water (3 mm high) to which about 10 MBq radioactive tracer has been added. The result depends mainly on the collimator, and will be worse if the resolution is better.

7.1.2 Whole body imaging

There is only one essential difference between planar imaging and whole body imaging: in whole body imaging the patient is continuously moving with respect to the camera. In some systems the patient bed is translated, in other systems the camera is moved, but the potential problems are the same. During whole body imaging a single large planar image is acquired (see figure 5.3), which is as wide as the crystal (let us call that the x-axis) but much longer (the y-axis). Consequently, the coordinates $(x_{\mathrm{wb}}, y_{\mathrm{wb}})$ in the large image are computed as

$$x_{\mathrm{wb}} = x \quad (7.12)$$
$$y_{\mathrm{wb}} = y + vt, \quad (7.13)$$

where (x, y) is the normal planar coordinate, v is the table speed and t is the time. The speed v must be expressed in pixels per s, while the motor of the bed is trying to move the bed at a predefined speed in cm per s. Consequently, whole body image quality will only be optimal if

- planar image quality is optimal, and
- the y-pixel size is exact, and
- the motor is working at the correct speed.

7.1.2.1 Bed motion

The bed motion can be measured directly to check if the true scan time is identical to the specified one.

7.1.2.2 Uniformity

If the table motion is not constant, the image of a uniform source will not be uniform. Of course, this uniformity test will also be affected if something is wrong with planar performance. Uniformity is not affected if the motion is wrong but constant, except possibly at the beginning and the end, depending on implementation details.

7.1.2.3 Pixel size, resolution

If there is something wrong with the conversion of table position to pixel coordinate, this is very likely to affect the pixel size and spatial resolution in the direction of table motion y. One can measure the pixel size directly by measuring two points with very different y_{wb}-coordinate. It is probably easier is to acquire a bar phantom or line phantom, which gives an immediate impression of the spatial resolution.

7.1.3 SPECT

7.1.3.1 Center-of-rotation

In SPECT, the gamma camera is used to acquire a set of sinograms, one for each slice. Sinogram data are two-dimensional, one coordinate is the angle, the other one the distance between the origin and the projection line. Consequently, a sinogram can only be correctly interpreted if we know which column in the image represents the projection line at zero distance. By definition, the is the point where the origin, the rotation center is projected.

For reasons of symmetry it is preferred (but not necessary) that the projection of the origin is the central column of the sinogram, and manufacturers attempt to obtain this. However, because of drift in the electronics, there can be a small offset. This is illustrated in figure 7.7. The leftmost images represent the sinogram of a point source and the corresponding reconstruction if everything works well. The rightmost images show what happens if there is a small offset between the projection of the origin and the center of the sinogram (the position where the origin is assumed to be during reconstruction, indicated with the arrows in the sinogram). For a point source in the center, the sinogram becomes inconsistent: the point is always at the right, no matter the viewing angle. As a result, artifacts result in filtered backprojection (the backprojection lines do not intersect where they should), which become more severe with increasing center-of-rotation error.

From the sinogram, one can easily compute the offset. Once the offset is known, we can correct it, either by shifting the sinogram to the left, or by telling the reconstruction algorithm where the true rotation center is. SPECT systems software provides automated procedure which compute the projection of the rotation axis from a point source measurement. Older cameras suffered a lot from this problem, newer cameras seem to be far more stable.

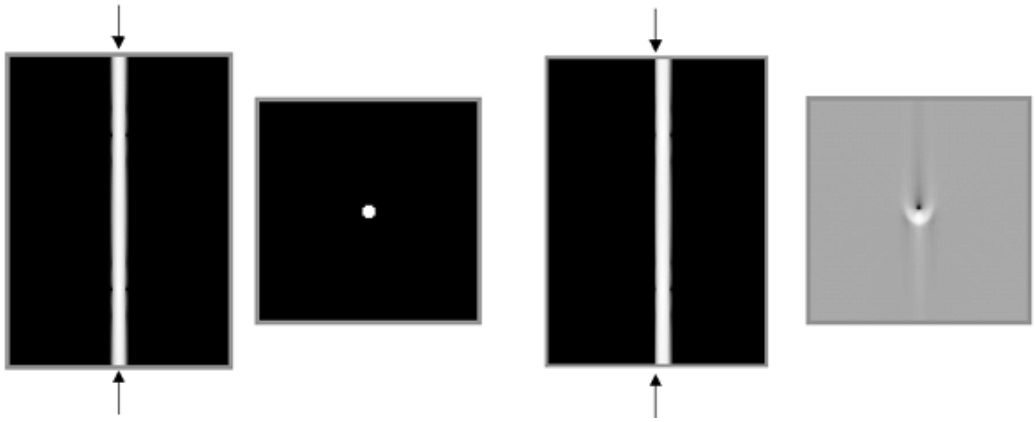

Figure 7.7: *Simulation of center of rotation error. Left: sinogram and reconstruction in absence of center of rotation error. Right: filtered backprojection with center of rotation error equal to the diameter of the point.*

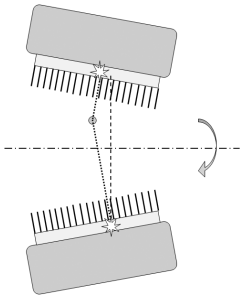

Figure 7.8: *If the gamma camera is not parallel to the rotation axis, a point apparently moves up and down the axis during rotation.*

7.1.3.2 Detector parallel to rotation axis

If the detector is not parallel to the rotation axis, the projection for different angles are not parallel. This gives rise to serious artifacts. Figure 7.8 shows that the projection of a point source seems to oscillate on the rotation axis while the camera rotates (except for points on the rotation axis).

The figure tells the way to detect the error: put a point source in the field of view, acquire a SPECT study and check if the point moves up and down in the projections (or disappears from one sinogram to show up in the next one). For a fixed angle, the amplitude of the apparent motion is proportional to the distance between the point and the rotation axis. Nicely centered point sources will detect nothing.

7.1.3.3 SPECT phantom

It is a good idea to scan a complex phantom every now and then, and compare the reconstruction to images obtained when the camera was known to be tuned optimally. Several companies offer polymethylmetacrylate (= perspex, Plexiglas) phantoms for this purpose. The typical phantom is a hollow cylinder which must be filled with water containing a radioactive tracer. Various inserts are provided, e.g. sets of bars of different diameter, allowing

to evaluate the spatial resolution. A portion of the cylinder is left free of inserts, allowing to check if a homogeneous volume is reconstructed into a homogeneous image. Almost any camera problem will lead to loss of resolution or loss of contrast.

When filling a phantom with "hot" water, do not count on diffusion of the tracer molecules to obtain a uniform distribution: it takes for ever before the phantom is truly uniform. The water must be stirred.

7.2 Positron emission tomograph

A PET-system is more expensive and more complicated than a gamma camera, because it contains more and faster electronics. However, in principle, it is easier to tune. The reason is that a multicrystal design has some robustness because of its modularity: if all modules work well, the whole system works well. A module contains only a few PMT's and a few crystals, and keeping such a small system well-tuned is "easier" than doing the same for a large single crystal detector.

7.2.1 Evolution of blank scan

Current PET-systems have computer controlled transmission sources: the computer can bring them in the field of view, and can have them retracted into a shielding container. When the transmission rod sources are extended they are continuously rotated near the perimeter of the field of view, such that all projection lines will be intersected.

A new blank scan is automatically acquired in the early morning, before working hours. This blank scan is a useful tool for performance monitoring, since all possible detector pairs contribute to it. The daily blank scan is compared to the reference blank scan, and significant deviations are automatically detected and reported, so drift in the characteristics is soon detected. If a problem is found it must be remedied. Often it is sufficient to redo calibration and normalization. Possibly a PMT or an electronic module is replaced, followed by calibration and normalization. When the system is back in optimal condition, a new reference blank scan is acquired.

7.2.2 Calibration

The term "calibration" refers to a complex automated procedure which

- tunes the PMT-gains to make their characteristics as similar as possible,
- determines the energy and position correction matrices for each detector module,
- determines differences in response time (a wire of 1 m adds 3 ns under optimal conditions) and uses that information to make the appropriate corrections, to ensure that the coincidence window (about 12 ns) meets its specifications.

The first two operations are performed with a calibrated phantom carefully positioned in the field of view, since 511 keV photons are needed to measure spectra and construct histograms as a function of position. The timing calibration is based on purely electronic procedures, without the phantom (since one cannot predict when the photons will arrive).

7.2.3 Normalization

"Normalization" in PET is identical to the "uniformity correction" of the gamma camera (PET-technology and gamma camera technology have been developed by different research groups and different companies, so the vocabulary is a bit different too). In a gamma camera, it is relatively easy to obtain an nearly homogeneous photon flux on the detector: a point source at large distance or a flat flood source are fine. In contrast, it is impossible to design a phantom that emits the same amount of photons along all possible projection lines for a ring detector system. Two approaches can be followed.

One approach is to slowly rotate a flat homogeneous phantom in the field of view, and only use measurements along projection lines (nearly) perpendicular to the phantom. It is time consuming, since most of the acquired data is ignored, but it provides a direct sensitivity measurement for all projection lines.

Alternatively, an indirect approach can be used. A sinogram is acquired for a large phantom in the center of the field of view. All individual detectors contribute to the measurement, and each detector is involved in multiple projection lines intersecting the phantom. However, many other detector pairs are not receiving photons at all. This measurement is sufficient to compute individual detector sensitivities. Sensitivity of a projection line can then be computed from the individual detector sensitivity and the known geometry. The result is a sinogram representing projection line sensitivities. This sinogram can be used to compute a correction for every sinogram pixel:

$$\text{normalization}(i) = \frac{\text{mean(sensitivity)}}{\text{sensitivity}(i)}. \tag{7.14}$$

A typical normalization sinogram is shown in figure 7.9. The figure seems a composition of lines: each line corresponds to a single detector (the location of the projection of the detector). As you can see, uniformity correction is far more important for PET than for the gamma camera. The reason is that the sensitivity of a single crystal depends on its position in the module, and that the geometry of the camera introduces sensitivity variations.

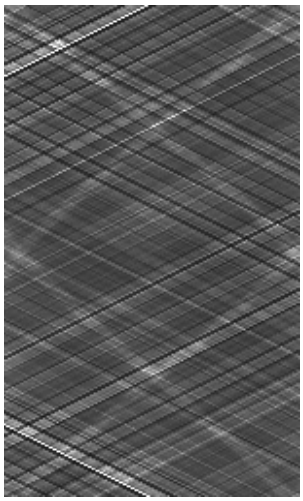

Figure 7.9: *PET normalization sinogram. The four lines bended lines of increased intensity are due to attenuation by transmission source hardware, located in the field of view.*

Chapter 8

Image analysis

In principle, PET and SPECT have the potential to provide quantitative images, that is pixel values can be expressed in Bq/ml. This is only possible if the reconstruction is based on a "sufficiently accurate" model, e.g. if attenuation is not taken into account, the images are definitely not quantitative. It is difficult to define "sufficiently accurate", the definition should certainly depend upon the application. Requirements for visual inspection are different from those for quantitative analysis.

In image analysis often *regions of interest* (ROI's) are used. A region of interest is a set of pixels which are supposed to belong together. Usually, an ROI is selected such that it groups pixels with (nearly) identical behavior, e.g. because they belong to the same part of the same organ. The pixels are grouped because the relative noise on the mean value of the ROI is lower than that on the individual pixels, so averaging over pixels results in strong noise suppression. ROI definition must be done carefully, since averaging over non-homogeneous regions leads to errors and artifacts! Ideally an ROI should be three-dimensional. However, mostly two dimensional ROI's defined in a single plane are used for practical reasons (manual ROI definition or two-dimensional image analysis software).

In this chapter only two methods of quantitative analysis will be discussed: standard uptake values (SUV) and the use of compartmental models. SUV's are simple to compute, which is an important advantage when the method has to be used routinely. In contrast, tracer kinetic analysis with compartmental models is often very time consuming, but it provides more quantitative information. In nuclear medicine it is common practice to study the kinetic behavior of the tracer, and many more analysis techniques exist. However, compartmental modeling is among the more complex ones, so if you understand that technique, learning the other techniques should not be too difficult. The book by Sorensen and Phelps [1] contains an excellent chapter on kinetic modeling and discusses several analysis techniques.

8.1 Standardized Uptake Value

The standardized uptake value provides a robust scale of tracer amounts. It is defined as:

$$\mathrm{SUV}_j = \frac{\text{tracer concentration in } j}{\text{average tracer concentration}} \tag{8.1}$$

$$= \frac{\text{tracer amount in Bq/g at pixel } j}{\text{total dose in Bq / total weight in g}} \tag{8.2}$$

To compute it, we must know the total dose administered to the patient. Since the total dose is measured prior to injection, and the image is produced after injection, we must correct for the decay of the tracer in between. Moreover, the tracer amounts are measured with different devices: the regional tracer concentration is measured with the SPECT or PET, the dose is measured with a dose calibrator. Therefor, the sensitivity of the tomograph must be determined. This is done by acquiring an image of a uniform phantom filled with a know tracer concentration. From the reconstructed phantom image we can compute a calibration factor which converts "reconstructed pixel values" into Bq/ml.

A SUV of 1 means that the tracer concentration in the ROI is identical to the average tracer concentration in the entire patient body. A SUV of 4 indicates markedly increased tracer uptake. The SUV-value is intended to be robust, independent from the administered tracer amount and the weight of the patient. However, it changes with time, since the tracer participates in a metabolic process. So SUVs can only be compared if they correspond to the same time after injection.

The SUV is a way to quantify the tracer concentration. But we don't really want to know that. The tracer was injected to study a metabolic process, so what we really want to quantify is the intensity of that process. The next section explains how this can be done. Many tracers have been designed to accumulate as a result of the metabolic process being studied. If the tracer is accumulated to high concentrations, many photons will be emitted resulting in a signal with good signal to noise ratio. In addition, although the tracer concentration is not nicely proportional to the metabolic activity, it is often an increasing function of that activity, so it still provides useful information.

8.2 Tracer kinetic modeling

8.2.1 Introduction

The evolution of the tracer concentration with time in a particular point (pixel) depends on the tracer and on the characteristics of the tissue in that point. Figure 8.1 shows the evolution of radioactive ammonia $^{13}NH_3$ (^{13}N is a positron emitter) in the heart region. This is a perfusion tracer: its concentration in tissue depends mainly on blood delivery to the cells. The tracer is injected intravenously, so it first shows up in the right atrium and ventricle. From there it goes to the lungs, and arrives in the left ventricle after another 20 s. After that, the tracer is gradually removed from the blood since it is accumulated in tissue. The myocardial wall is strongly perfused, after 20 min accumulation in the left ventricular wall is very high, and even the thin right ventricular wall is clearly visible.

The most important factor determining the dynamic behavior of ammonia is blood flow. However, the ammonia concentration is not *proportional* to blood flow. To quantify the blood flow in ml blood per g tissue and per s, the flow must be computed from dynamic behavior of the tracer. To do that, we need a mathematical model that describes the most important features of the metabolism for this particular tracer. Since different tracers trace different metabolic processes, a different model is required for each tracer.

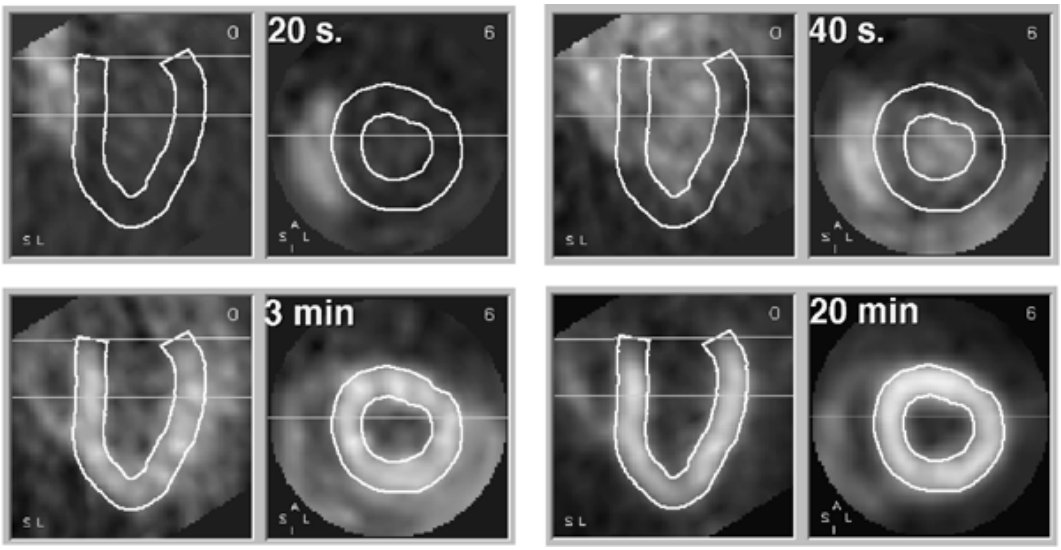

Figure 8.1: *Four samples from a dynamic study. Each sample shows a short axis and long axis slice through the heart. The wall of the left ventricle is delineated. 20 s after injection, the tracer is in the right ventricle. At 40 s it arrives in the left ventricle. At 3 min, the tracer is accumulating in the ventricular wall. At 20 min, the tracer concentration in the wall is increased, resulting in better image quality.*

8.2.2 The compartmental model

8.2.2.1 The compartments

In this section, the three compartment model is described. It is a relatively general model and can used for a few different tracers. We will focus on the tracer ^{18}F-fluorodeoxyglucose (FDG), which is a glucose analog. "Analog" means that it is *no* glucose, but that it is sufficiently similar to follow, to some extent, the same metabolic pathway as glucose. In fact, it is a better tracer than radioactive glucose, because it is trapped in the cell, while glucose is not. When glucose enters the cell, it is metabolized and the metabolites may escape from the cell. As a result, radioactive glucose will never give a strong signal. In contrast, FDG is not completely metabolized (because of the missing oxide), and the radioactive ^{18}F atom stays in the cell. If the cells have a high metabolism, a lot of tracer will get accumulated resulting in a strong signal (many photons will be emitted from such a region, so the signal to noise ratio in the reconstructed image will be good).

Figure 8.2 shows the three compartmental model and its relation to the measured concentrations. The first compartment represents the radioactive tracer molecule present in blood plasma. The second compartment represents the unmetabolized tracer in the "extravascular space", that is, everywhere except in the blood. The third compartment represents the radioactive isotope after metabolization (so it may now be part of a very different molecule). In some cases, the compartments correspond nicely to different structures, but in other cases the compartments are abstractions. E.g., the second two compartments may be present in a single cell. We will denote the blood plasma compartment with index **P**, the extravascular compartment with index **E** and the third compartment with index **M**.

These compartments are an acceptable simplified model for the different stages in the FDG

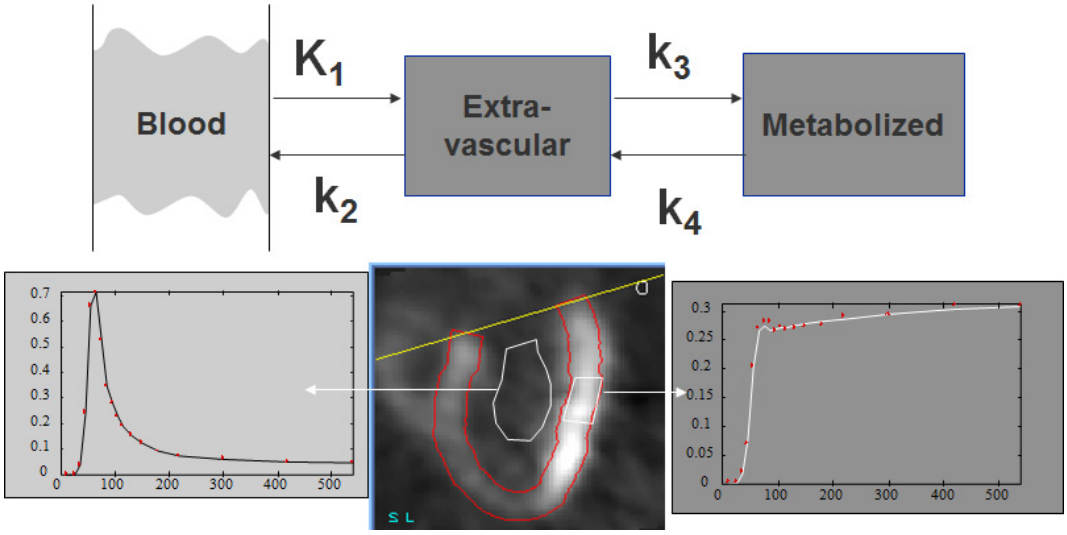

Figure 8.2: *Three compartment model, and corresponding regions in a cardiac study. The first compartment represents plasma concentration (blood pool ROI), the second and third compartment represent the extavascular tracer in original and metabolized form (tissue ROI). For the tissue curve, both the measured points and the fitted curve are shown.*

and glucose pathways. A model for another tracer may need fewer or more compartments.

8.2.2.2 The rate constants

The amount of tracer in a compartment can be specified in several ways: number of molecules, mole, gram or Bq. All these numbers are directly proportional to the absolute number of molecules. If we study a single gram of tissue, the amounts can be expressed in Bq/g, which is the unit of concentration. Remark, though, that these are not tracer concentrations of the compartments, since the gram tissue contains several compartments.

The compartments may exchange tracer molecules with other compartments. It is assumed that this exchange can be described with simple first order rate constants. This means that the amount of tracer traveling away from a compartment is proportional to the amount of tracer in that compartment. The constant of proportionality is called a *rate constant*. For the three compartment model, there are two different types of rate constants. Constant K_1 is a bit different from k_2, k_3 and k_4 because the first compartment is a bit different from the other two (fig. 8.2).

The amount of tracer going from the blood plasma compartment to the extravascular compartment is

$$\text{Tracer}_{P\to E} = K_1 C_P \tag{8.3}$$

where C_P is the plasma tracer concentration (units: Bq/ml). K_1 is the product of two factors. The first one is the blood flow F, the amount of blood supplied in every second to the gram tissue. F has units ml/(s g). The second factor is the extraction fraction f, which specifies which fraction of the tracer passing by in P is exchanged with the second compartment E. Since it is a fraction, f has no unit. For some tracers, f is almost 1 (e.g. for radioactive water). For other tracers it is 0, which means that the tracer stays in the blood. For FDG it is in between. Consequently, the product $K_1 C_P = F f C_P$ has units Bq/(s g) and tells how

many Bq are extracted from the blood per second and per gram tissue. C_P is a function of the time, K_1 is not. Both K_1 and C_P are also a function of position: the situation will be different in different tissue types.

The amount of tracer going from E to M equals:

$$\text{Tracer}_{E\to M} = k_3 C_E(t), \tag{8.4}$$

where C_E is the total amount of tracer in compartment E per gram tissue (units: Bq/g). Constant k_3 tells which fraction of the available tracer is metabolized in every second, so the unit of k_3 is 1/s. The physical meaning of k_2 and k_4 is similar: they specify the fractional transfer per second.

8.2.2.3 The target molecule

The tracer is injected to study the metabolism of a particular molecule. It is assumed that the metabolic process being studied is constant during the measurement, and that the target molecule (glucose in our example) has reached a steady state situation. Steady state means that a dynamic equilibrium has been reached: all concentrations remain constant. Steady state can only be reached with well-designed feedback systems (poor feedback systems oscillate), but it is reasonable to assume that this is the case for metabolic processes in living creatures.

Since FDG and glucose are not identical, their rate constants are not identical. The glucose rate constants will be labeled with the letter g. The glucose and FDG amounts are definitely different: glucose is abundantly present and is in steady state condition. FDG is present in extremely low concentrations (pmol) and has not reached steady state since the images are acquired immediately after injection.

The plasma concentration C_P^g is supposed to be constant. We can measure it by determining the glucose concentration in the plasma from a venous blood sample. The extravascular glucose amount C_E^g is supposed to be constant as well, so the input must equal the output. For glucose, k_4^g is very small. Indeed, k_4^g corresponds to reversal of the initiated metabolization, which happens only rarely. Setting k_4^g to zero we have

$$\frac{dC_E^g}{dt} = 0 = K_1^g C_P^g - (k_2^g + k_3^g) C_E^g. \tag{8.5}$$

Thus, we can compute the unknown tracer amount C_E^g from the known plasma concentration C_P^g:

$$C_E^g = \frac{K_1^g}{k_2^g + k_3^g} C_P^g. \tag{8.6}$$

The glucose metabolization in compartment M is proportional to the glucose transport from E to M. Of course, the glucose metabolites may be transported back to the blood, but we don't care. We are only interested in glucose. Since it ceases to exist after transport to compartment M we ignore all further steps in the metabolic pathway. In our compartment model, it is as if glucose is accumulated in the metabolites compartment. This virtual accumulation rate is the metabolization rate we want to find:

$$\frac{dC_M^g}{dt} = k_3^g C_E^g = \frac{K_1^g k_3^g}{k_2^g + k_3^g} C_P^g. \tag{8.7}$$

$$= \bar{K}^g C_P^g. \qquad \text{definition of } \bar{K}^g \tag{8.8}$$

So if we can find the values of the rate constants, we can compute the glucose metabolization rate. As mentioned before, we cannot compute them via the tracer, since it has different rate constants. However, it can be shown that, due to its similarity, the trapping of FDG is *proportional* to the glucose metabolic rate. The constant of proportionality depends on the tissue type (difference in affinity for both molecules), but not on the tracer concentration. The constant of proportionality is usually called "the lumped constant", because careful theoretical analysis shows that it is a combination of several constants. So the lumped constant LC is:

$$LC = \frac{\frac{K_1 k_3}{k_2+k_3}}{\frac{K_1^g k_3^g}{k_2^g+k_3^g}} = \frac{\bar{K}}{\bar{K}^g} \tag{8.9}$$

For the brain (which gets almost all its energy from glucose) and for some other organs, the lumped constant in humans has been measured. Note that if we had used radioactive glucose, the tracer and target molecules would have had identical rate constants and the lumped constant would have been 1.

8.2.2.4 The tracer

Problem statement

Because the lumped constant is known, we can compute the glucose metabolization rate from the FDG trapping rate. To compute the trapping rate, we must know the FDG rate constants. Since the tracer is not in steady state, the equations will be a bit more difficult than for the target molecule. We can easily derive differential equations for the concentration changes in the second and third compartment:

$$\frac{dC_E(t)}{dt} = K_1 C_P(t) - (k_2 + k_3) C_E(t) \tag{8.10}$$

$$\frac{dC_M(t)}{dt} = k_3 C_E(t) \tag{8.11}$$

For a cardiac study, we can derive the tracer concentration $C_P(t)$ in the blood from the pixel values in the center of the left ventricle or atrium. If the heart is not in the field of view, we can still determine $C_P(t)$ by measuring the tracer concentrations in blood samples withdrawn at regular time intervals. As with the SUV computations, this requires cross-calibration of the plasma counter to the PET camera.

The compartments E and M can only be separated with subcellular resolution, so the PET always measures the sum of both amounts, which we will call $C_I(t)$:

$$C_I(t) = C_E(t) + C_M(t). \tag{8.12}$$

Consequently, we must combine the equations (8.10) and (8.11) in order to write $C_I(t)$ as a function of $C_P(t)$ and the rate constants. This is the operational equation. Since $C_I(t)$ and $C_P(t)$ are known, the only remaining unknown variables will be the rate constants, which are obtained by solving the operational equation.

Deriving the operational equation

To deal with differential equations, the Laplace transform is a valuable tool. Appendix 10.7 gives the definition and a short table of the features we need for the problem at hand. The

strength of the Laplace transform is that derivatives and integrals with respect to t become simple functions of s. After transformation, elimination of variables is easy. The result is then back-transformed to the time domain. Laplace transform of (8.10) and (8.11) results in

$$\begin{aligned} sc_E(s) &= K_1 c_P(s) - (k_2 + k_3) c_E(s) && (8.13) \\ sc_M(s) &= k_3 c_E(s) && (8.14) \end{aligned}$$

where we have assumed that at time $t = 0$ (time of injection) all tracer amounts are zero. From (8.13) we find $c_E(s)$ as a function of $c_P(s)$. Inserting in (8.14) produces $c_M(s)$ as a function of $c_P(s)$.

$$\begin{aligned} c_E(s) &= \frac{K_1}{s + k_2 + k_3} c_P(s) && (8.15) \\ c_M(s) &= \frac{K_1 k_3}{s(s + k_2 + k_3)} c_P(s) && (8.16) \\ c_I(s) &= c_E(s) + c_M(s) && (8.17) \\ &= \left(\frac{K_1}{s + k_2 + k_3} + \frac{K_1 k_3}{s(s + k_2 + k_3)} \right) c_P(s) && (8.18) \end{aligned}$$

The two factors in s can be split from the denominator using the equation

$$\frac{a}{x(x+b)} = \frac{a}{b} \left(\frac{1}{x} - \frac{1}{x+b} \right) \tag{8.19}$$

Applying this to (8.18) and rearranging a bit yields:

$$c_I(s) = \frac{K_1 k_2}{(k_2 + k_3)} \frac{c_P(s)}{(s + k_2 + k_3)} + \frac{K_1 k_3}{(k_2 + k_3)} \frac{c_P(s)}{s} \tag{8.20}$$

Applying the inverse Laplace transform is now straightforward (see appendix 10.7) and produces the operational equation:

$$C_I(t) = \frac{K_1 k_2}{k_2 + k_3} \int_0^t C_P(u) e^{-(k_2+k_3)(t-u)} du + \frac{K_1 k_3}{k_2 + k_3} \int_0^t C_P(u) du. \tag{8.21}$$

Figure 8.3 plots C_I and the two terms of equation (8.21) for the case when $C_P(t)$ is a step function. $C_P(t)$ is never a step function, but the plot provides interesting information. The first term of (8.21) represents tracer molecules that leave the vascular space, stay a while in compartment E and then return back to the blood. As soon as the tracer is present in the blood, this component starts to grow until it reaches a maximum. When $C_P(t)$ becomes zero again, the component gradually decreases towards zero. This first term follows the input, but with some delay (CI_1(t) in fig. 8.3).

The second term of (8.21) represent tracer molecules that enter compartment E and will never leave (CI_2(t) in fig. 8.3). Eventually, they will be trapped in compartment M. Note that the first term is not equal to but smaller than $C_E(t)$. The reason is that part of the molecules in E will not return to the blood but end up in M. It is easy compute which fraction of $C_E(t)$ is described by the first term of (8.21). (The rest of $C_E(t)$ and $C_M(t)$ correspond to the second term of (8.21).) This is left as an exercise to the reader.

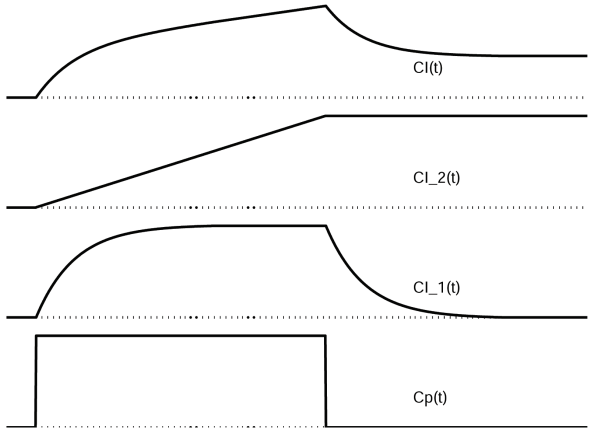

Figure 8.3: *The tracer amount $C_I(t)$ and its two terms when $C_P(t)$ is a step function (equation (8.21)).*

Computing the rate constants with non-linear regression

At this point, we have the operational equation relating $C_I(t)$ to $C_p(t)$ and the rate constants. We also know $C_p(t)$ and $C_I(t)$, at least in several sample points (dynamic studies have typically 20 to 40 frames). Every sample point is an equation, so we actually have a few tens of equations and 3 unknowns. Because of noise and the fact that the operational equation is only an approximation, there is probably no exact solution. The problem is very similar to the reconstruction problem, which has been solved with the maximum likelihood approach. The same will be done here. If the likelihood is assumed to be Gaussian, maximum likelihood becomes weighted least squares.

With this approach, we start with an arbitrary set of rate constants. It is recommended to start close to the solution if possible, because the likelihood function may have local maxima. Typically the rate constants obtained in healthy volunteers are used as a start. With these rate constants and the known input function $C_p(t)$, we can compute the expected value of $C_I(t)$. The computed curve will be different from the measured one. Based on this difference, the non-linear regression algorithm will improve the values of the rate constants, until the sum of weighted squared differences is minimal. It is always useful to plot both the measured and computed curves $C_I(t)$ to verify that the fit has succeeded, since there is small chance that the solution corresponds to an unacceptable local minima. In that case, the process must be repeated, starting from a different set of rate constant values. The tissue curve in fig. 8.2 is the result of non-linear regression. The fit was successful, since the curve is close to the measured values.

The glucose consumption can now be computed as

$$\text{glucose consumption} = \frac{1}{\text{LC}} \frac{K_1 k_3}{k_2 + k_3} C_P^g = \frac{\bar{K}}{\text{LC}} C_P^g \tag{8.22}$$

Non-linear regression programs often provide and estimate of the confidence intervals or standard deviations on the fitted parameters. These can be very informative. Depending on the shape of the curves, the noise on the data and the mathematics on the model, the accuracy of the fitted parameters can be very poor. However, the errors on the parameters are correlated such that the accuracy on $\bar{K}$ is better than the accuracy on the individual rate constants.

Computing the trapping rate with linear regression

By introducing a small approximation, the computation of the glucose consumption can be simplified. Figure 8.2 shows a typical blood function. The last part of the curve is always very smooth. As a result, the first term of (8.21) nicely follows the shape of $C_p(t)$. Stated otherwise, $C_p(u)$ changes very little over the range where $e^{-(k_2+k_3)(t-u)}$ is significantly different from zero. Thus, we can put $C_p(t)$ in front of the integral sign. Since t is large relative to the decay time of the exponential, we can set t to ∞:

$$\begin{aligned}
\int_0^t C_P(u)e^{-(k_2+k_3)(t-u)}du &\simeq C_P(t)\int_0^t e^{-(k_2+k_3)(t-u)}du && (8.23)\\
&= C_P(t)\int_0^t e^{-(k_2+k_3)u}du && (8.24)\\
&\simeq C_P(t)\int_0^\infty e^{-(k_2+k_3)u}du && (8.25)\\
&= \frac{C_p(t)}{k_2+k_3} && (8.26)
\end{aligned}$$

The operational equation then becomes:

$$C_I(t) \simeq \frac{K_1k_3}{k_2+k_3}\int_0^t C_P(u)du + \frac{K_1k_2}{(k_2+k_3)^2}C_P(t). \tag{8.27}$$

Now both sides are divided by $C_p(t)$:

$$\frac{C_I(t)}{C_P(t)} \simeq \frac{K_1k_3}{k_2+k_3}\frac{\int_0^t C_P(u)du}{C_P(t)} + \frac{K_1k_2}{(k_2+k_3)^2}. \tag{8.28}$$

Equation (8.28) says that $C_I(t)/C_P(t)$ is a linear function of $\int_0^t C_P(u)du/C_P(t)$, at least for large values of t. We can ignore the constants, all we need is the slope of the straight curve. This can be obtained with simple linear regression. For linear regression no iterations are required, there is a closed form expression, so this solution is orders of magnitudes faster to compute than the previous one.

The integral $\int_0^t C_P(u)du/C_P(t)$ has the unit of time. If $C_P(t)$ would be a constant, the integral simply equals t. It can be regarded as a correction, required because $C_P(t)$ is not constant but slowly varying. As shown in figure 8.3, when $C_P(t)$ is constant, $C_I(t)$ has a linear and a constant term. Equation (8.21) confirms that the slope of the linear term is indeed $\bar{K}$. A potential disadvantage is that the values of the rate constants are not computed.

8.3 Image quality

It is extremely difficult to give a useful definition of image quality. As a result, it is even more difficult to measure it. Consequently, often debatable measures of image quality are being used. This is probably unavoidable, but it is good to be fully aware about the limitations.

8.3.1 Subjective evaluation

A very bad but very popular way to assess the performance of some new method is to display the image produced by the new method together with the image produced in the classical

way, and see if the new image is "better". In many cases, you cannot see if it is better. You can see that you like it better for some reason, but that does not guarantee that it will lead to an improvement in the process (e.g. making a diagnosis) to which the image is supposed to contribute.

E.g. a study has been carried out to determine the effect of 2D versus 3D PET imaging on the diagnosis of a particular disease, and for a particular PET system. In addition, the physicians were asked to tell what images they preferred. The physicians preferred the 3D images because they look nicer, but their diagnosis was statistically significantly better on the 2D images (because scatter contribution was lower).

It is not forbidden to look at an image, but it is important not to jump to a conclusion.

8.3.2 Task dependent evaluation

The best way to find out if an image is good, is to use it as planned and check if the results are good. Consequently, if a new image generation or processing technique is introduced, it has to be compared very carefully to the classical method on a number of clinical cases. Evaluation must be done blindly: if the observer remembers the image from the first method when scoring the image from the second method, the score is no longer objective. If possible the observer should not even be aware of the method, in order to exclude the influence of possible prejudice about the methods.

8.3.3 Continuous and digital

Intuitively, the best image is the one closest to the truth. But in emission tomography, the truth is a continuous tracer distribution, while the image is digital. It is not always straightforward to define how a portion of a continuous curve can be best approximated as a single value. The problem becomes particularly difficult if the images to be compared use a different sampling grid (shifted points or different sampling density). So if possible, make sure that the sampling is identical.

8.3.4 Bias and variance

Assume that we have a reference image e.g. in a simulation experiment. In this case, we know the true image, and in most cases it is even digital. Then we can compare the difference between the true image and the image to be evaluated. A popular approach is to compute the mean squared difference:

$$\text{mean square difference} = \frac{1}{J}\sum_{j=1}^{J}(\lambda_j - r_j)^2, \tag{8.29}$$

where λ_j and r_j are the image and the reference image respectively. This approach has two problems. First, it assumes that all pixels are equally important, which is almost never true. Second, it combines systematic (bias) and random (variance) deviations. It is better to separate the two, because they behave very differently. This is illustrated in figure 8.4. Suppose that a block wave was measured with two different methods A and B, producing the noisy curves shown at the top row. Measurement A is noisier, and its sum of squared differences with the true wave is twice as large as that of measurement B. If we know that we are measuring a block wave, we know that a bit of smoothing is probably useful. The bottom

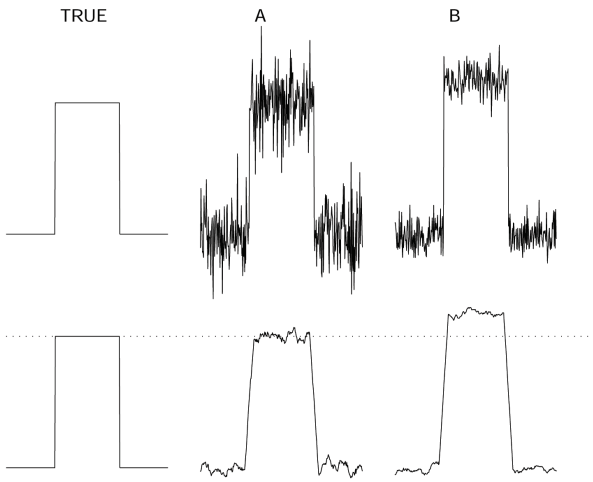

Figure 8.4: *Top: a simulated true block wave and two measurements A and B with different noise distributions. Bottom: the true block wave and the same measurements, smoothed with a simple rectangular smoothing kernel.*

row shows the result of smoothing. Now measurement A is clearly superior. The reason is that the error in the measurement contains both bias and variance. Smoothing reduces the variance, but increases the bias. In this example the true wave is smooth, so variance is strongly reduced while the bias only increases near the edges of the wave. If we keep on smoothing, the entire wave will converge to its mean value; then there is no variance but huge bias. Bias and variance of a sample (pixel) j can be defined as follows:

$$\text{bias}_j = E(\lambda_j - r_j) \tag{8.30}$$

$$\text{variance}_j = E\left((\lambda_j - E(\lambda_j))^2\right), \tag{8.31}$$

where $E(x)$ is the expectation of x. Variance can be directly computed from repeated independent measurements. If the true data happen to be smooth and if the measurement has good resolution, neighboring samples can be regarded as independent measurements. Bias can only be computed if the true value is known.

In many cases, there is the possibility to trade in variance for bias by smoothing or imposing some constraints. Consequently, if images are compared, bias and variance must be separated. If an image has better bias for the same variance, it is probably a "better" image. If an image has larger bias and lower variance when compared to another image, the comparison is meaningless.

8.3.5 Evaluating a new algorithm

A good way to evaluate a new image processing algorithm (e.g. an algorithm for image reconstruction, for image segmentation, for registration of images from different modalities) is to apply the following sequence:

1. evaluation on computed, simulated data

2. evaluation on phantom data

3. (evaluation on animal experiments)

4. evaluation on patient data

This sequence is in order of increasing complexity and decreasing controllability. Tests on patient data are required to show that the method can be used. However, if such a test fails it is usually very difficult to find out why. To find the problem, simple and controllable data are required. Moreover, since the true solution is often not known in the case of patient data, it is possible that failure of the method remains undetected. Consequently, there is no gain in trying to skip one or a few stages, and with a bit of bad luck it can have serious consequences.

Evaluation on simulation has the following important advantages:

- The truth is known, comparing the result to the true answer is simple. This approach is also very useful for finding bugs in the algorithm or its implementation.
- Data can be generated in large numbers, sampling a predefined distribution. This enables direct quantitative analysis of bias and variance.
- Complexity can be gradually increased by making the simulations more realistic, to analyze the response to various physical phenomena (noise, attenuation, scatter, patient motion ...).
- A nice thing about emission tomography is that it is relatively easy to make realistic simulations. In addition, many research groups are willing to share simulation code.
- It is possible to produce simulations which are sufficiently realistic to have them diagnosed by the nuclear medicine physicians. Since the correct diagnosis is known, this allows evaluation of the effect of the new method on the final diagnosis.

When the method survives complex simulations it is time to do phantom experiments. Phantom experiments are useful because the true system is always different from even a very realistic simulation. If the simulation phase has been done carefully, phantom experiments are not likely to cause much trouble.

A possible intermediate stage is the use of animal experiments, which can be required for the evaluation of very delicate image processing techniques (e.g. preparing stereotactic operations). Since the animal can be sacrificed, it is possible, at least to some extent, to figure out what result the method should have produced.

The final stage is the evaluation on patient data, and comparing the output of the new method to that of the classical method. As mentioned before, not all failures will be detected since the correct answer may not be known.

Chapter 9

Biological effects

For this subject, we refer to chapter 10 "Internal Radiation Dosimetry" in the book by Sorensen and Phelps [1], which shows how the radiation dose in every organ can be estimated if the dynamic behavior of the tracer is known.

The radiation dose is expressed in gray (Gy). A dose of 1 gray means that 1 J (joule) energy has been deposited per kg absorber:

$$1Gy = \frac{1J}{kg}. \tag{9.1}$$

For dosimetry, it is important to quantify how "bad" this radiation is. It turns out that 1 J deposited by neutrons does more damage than the same energy deposited by photons. To take this effect into account, a quality factor Q is introduced. Multiplication with the quality factor converts the dose into the dose equivalent. Quality factors are given in table 9.1.

The equivalent dose is expressed in Sv (sievert). Since $Q = 1$ for photons and positrons, we have 1 mSv per mGy in diagnostic nuclear medicine. The natural background radiation is about 2 mSv per year.

The older units for radiation dose and equivalent dose are rad and rem:

$$1 \text{ Gy} = 100 \text{ rad} \tag{9.2}$$

$$1 \text{ Sv} = 100 \text{ rem} \tag{9.3}$$

When the dose to every organ is computed, one can in addition compute an "effective dose", which is a weighted sum of organ doses. The weights are introduced because damage in one organ is more dangerous than damage in another organ. The most sensitive organs are the gonads (weight about 0.25), the breast, the bone marrow and the lungs (weight about

Table 9.1: Quality factor converting dose in equivalent dose.

Radiation	Q
X-ray, γ-ray, electrons, positrons	1
neutrons, protons	10
α-particals	20

0.15), the thyroid and the bones (weight about 0.05). The sum of all the weights is 1. The weighted average produces a single value in mSv. The risk of death due to tumor induction is about 5% per "effective Sv" according to report ICRP-60 (International Commission on Radiological Protection (http://www.icrp.org)), but it is of course very dependent upon age (e.g. 14% for children under 10). Research in this field is not finished and the tables and weighting coefficients are adapted every now and then.

The text uses the old units (rad, Ci and erg). To figure out what this means in SI units (Gy, Bq and J), you need to know that the charge of an electron is 1.6×10^{-19} Coulomb. Recall that Coulomb $\times$ Volt = Joule, and that 1 Gy = 1 J per kg. Consequently, if we have 1 MeV per disintegration, we can compute:

$$1 \text{ MeV} = 1.6 \times 10^{-13} \text{ J} \tag{9.4}$$

$$= 1.6 \times 10^{-13} \text{ J} \times \frac{10^6}{\text{MBq s}} \times \frac{3600 \text{ s}}{\text{h}} \tag{9.5}$$

$$= 5.76 \times 10^{-4} \frac{\text{Gy kg}}{\text{MBq h}} \tag{9.6}$$

This can be converted to the old units as follows:

$$1 \text{ MeV} = 5.76 \times 10^{-4} \frac{\text{Gy kg}}{\text{MBq h}} \times 100 \frac{\text{rad}}{\text{Gy}} \times 10^{-3} \frac{\text{g}}{\text{kg}} \times 37 \times 10^3 \frac{\text{MBq}}{\mu\text{Ci}} \tag{9.7}$$

$$= 2.13 \frac{\text{rad g}}{\mu\text{Ci h}} \tag{9.8}$$

The Society of Nuclear Medicine distributes MIRD software to do the (very lengthy) dose calculations described in Sorensen's book, so you will never have to do it manually.

Chapter 10

Appendix

10.1 Poisson noise

Assume that the expected amount of measured photons per unit time equals r. Let us now divide this unit of time in a k time intervals. If k is sufficiently large, the time intervals are small, so the probability of detecting a photon in such an interval is small, and the probability of detecting two or more is negligible. Consequently, we can assume that in every possible measurement, only zero or one photon is detected in every interval. The probability of detecting one photon in an interval is then r/k. A measurement of n photons must consist of n intervals with one photon, and $k-n$ intervals with no photons. The probability of such a measurement is:

$$p_r(n) = \left(\frac{r}{k}\right)^n \left(1 - \frac{r}{k}\right)^{k-n} \frac{k!}{n!(k-n)!} \tag{10.1}$$

The first factor is the probability of detecting n photons in n intervals. The second factor is the probability of detecting no photons in $n-k$ intervals. The third factor is the amount of ways that these n successful and $n-k$ unsuccessful intervals can be combined in different measurements.

As mentioned above, equation (10.1) becomes better when k is larger. To make it exact, we simply have to let k go to infinity.

$$\lim_{k\to\infty} p_r(n) = \frac{r^n}{n!} \lim_{k\to\infty} \left(1 - \frac{r}{k}\right)^k \tag{10.2}$$

It turns out that computing the limit of the logarithm is easier, so we have

$$\begin{aligned} \lim_{k\to\infty} \left(1 - \frac{r}{k}\right)^k &= \exp\left(\lim_{k\to\infty} \frac{\ln(k-r) - ln(k)}{1/k}\right) \\ &= \exp\left(\lim_{k\to\infty} \frac{1/(k-r) - 1/k}{-1/k^2}\right) \\ &= \exp(-r) \end{aligned} \tag{10.3}$$

So it follows that

$$\lim_{k\to\infty} p_r(n) = \frac{e^{-r} r^n}{n!}. \tag{10.4}$$

10.2 Convolution

In section 4.2.1, the collimator point spread function (PSF) was computed. The collimator PSF tells us which image is obtained for a point source at distance H from the collimator. What happens if two point sources are positioned in front of the camera, both at the same distance H? Since the sources and the photons don't interact with each other, all what was said above still applies, for each of the sources. The resulting image will consist of two PSFs, each centered at the detector point closest to the point source. Where the PSFs overlap, they must be added, since the detector area in the overlap region gets photons from both sources. The same is true for three, four, or one million point sources, all located at the same distance H from the collimator. To predict the image for a set of one million point sources, simply calculate the corresponding PSFs centered at the corresponding positions on the detector, and sum everything.

The usual mathematical description of this can be considered as a two step approach:

1. Assume that the system is perfect: the image of a point source is a point, located on the perpendicular projection line through the point source. Mathematicians would call that "point" in the image a "Dirac impulse". The image of two or more point sources contains simply two or more Dirac impulses, located at corresponding projection lines. Let $f(x, y)$ represent the image value at position (x, y). This image can be regarded as the sum of an infinite number of Dirac impulses $\delta(x, y)$, one at every location (x, y):

$$\begin{aligned} f(x,y) &= (f \otimes \delta)(x,y) && (10.5)\\ &= \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, f(u,v)\, \delta(x-u, y-v) && (10.6) \end{aligned}$$

2. Correction of the first step: the expected image of a point is not a Dirac impulse, but a PSF. Therefore, replace each of the Dirac impulses in the image by the corresponding PSF, centered at the position of the Dirac impulse.

$$\begin{aligned} g(x,y) &= (f \otimes \mathrm{PSF})(x,y) && (10.7)\\ &= \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, f(u,v)\, \mathrm{PSF}(x-u, y-v) && (10.8) \end{aligned}$$

The second operation is called the *convolution* operation. Assume that a complex flat (e.g. a inhomogeneous radioactive sheet) tracer distribution would be put in front of the camera, parallel to the collimator (at distance H). What image will be obtained? Regard the distribution as consisting of an infinite number of point sources. This does not change the distribution, it only changes the way you look at it. Project all of these sources along an ideal perpendicular projection line into the image. You will now obtain an image consisting of an infinite number of Dirac impulses. Replace each of these impulses with the PSF and sum the overlapping values to obtain the expected image.

If the distance to the collimator is not the same for every point source, then things get more complicated. Indeed, the convolution operator assumes that the PSF is the same for all point sources. Therefore, to calculate the expected image, the previous procedure has to be applied individually for every distance, using the corresponding distance dependent PSF. Finally, all convolved images have to be summed to yield the expected image.

10.3 Combining resolution effects: convolution of two Gaussians

Very often, the PSF can be well approximated as a Gaussian. This fact comes in handy if we want to combine two PSFs. For example: what is the combined effect of the intrinsic resolution (PSF of scintillation detection) and the collimator resolution (collimator PSF)?

How can two PSFs be combined? The solution is given in appendix 10.2: one of the PSFs is regarded as a collection of Dirac impulses. The second PSF must be applied to each of these pulses. So we must compute the convolution of both PSFs. This appendix shows that if both are approximately Gaussian, the convolution is easy to compute.

Let us represent the first and second PSFs as follows:

$$F_1(x) = A\exp\left(-\frac{x^2}{a^2}\right) \quad \text{and} \quad F_2(x) = B\exp\left(-\frac{x^2}{b^2}\right) \tag{10.9}$$

Thus, $\sigma_1 = a/\sqrt{2}$ and $A = 1/(\sqrt{2\pi}\sigma_1)$, and similar for the other PSF. The convolution is then written as:

$$\begin{aligned}(F_1 \otimes F_2)(x) &= AB\int_{-\infty}^{\infty}\exp\left(-\frac{u^2}{a^2} - \frac{(x-u)^2}{b^2}\right)du &(10.10)\\ &= AB\int_{-\infty}^{\infty}\exp\left(-mu^2 + \frac{2xu}{b^2} - \frac{x^2}{b^2}\right)du &(10.11)\\ m &= \left(\frac{1}{a^2} + \frac{1}{b^2}\right) &(10.12)\end{aligned}$$

The exponentiation contains a term in u^2 and a term in u. We can get rid of the u by requiring that both terms result from squaring something of the form $(u+C)^2$, and rewriting everything as $(u+C)^2 - C^2$. The C^2 is not a function of u, so it can be put in front of the integral sign.

$$\begin{aligned}(F_1 \otimes F_2)(x) &= AB\int_{-\infty}^{\infty}\exp\left(-\left(\sqrt{m}u - \frac{x}{b^2\sqrt{m}}\right)^2 + \frac{x^2}{b^4 m} - \frac{x^2}{b^2}\right)du &(10.13)\\ &= AB\exp\left(\frac{x^2}{b^4 m} - \frac{x^2}{b^2}\right)\int_{-\infty}^{\infty}\exp\left(-\left(\sqrt{m}u - \frac{x}{b^2\sqrt{m}}\right)^2\right)du &(10.14)\end{aligned}$$

The integrand is a Gaussian. The center is a function of x, but the standard deviation is not. The integral from $-\infty$ to ∞ of a Gaussian is a finite value, only dependent on its standard deviation. Consequently, the integral is not a function of x. Working out the factor in front of the integral sign and combining all constants in a new constant D, we obtain

$$(F_1 \otimes F_2)(x) = D\exp\left(-\frac{x^2}{a^2+b^2}\right) \tag{10.15}$$

So the convolution of two Gaussians is again a Gaussian. The variance of the resulting Gaussian is simply the sum of the input variances (by definition, the variance is the square of the standard deviation).

The FWHM of a Gaussian is proportional to the standard deviation, so we obtain a very simple expression to compute the FWHM resulting from the convolution of multiple PSFs:

$$\text{FWHM}^2 = \text{FWHM}_1^2 + \text{FWHM}_2^2 + \ldots + \text{FWHM}_n^2 \tag{10.16}$$

10.4 Error propagation

The mean and the variance of a distribution a are defined as:

$$\text{mean}(a) = \overline{a} = E(a) = \frac{1}{N}\sum_{i=1}^{N} a_i \tag{10.17}$$

$$\text{variance}(a) = \sigma_a^2 = E\left[(a-\overline{a})^2\right] = \frac{1}{N}\sum_{i=1}^{N}(a_i-\overline{a})^2, \tag{10.18}$$

where E is the expectation, a_i is a sample from the distribution and the summation is over the entire distribution. By definition, σ_a is the standard deviation. When computations are performed on samples from a distribution (e.g. Poisson variables) the noise propagates into the result. Consequently, one can regard the result also as a sample from some distribution which can be characterized by its (usually unknown) mean value and its variance or standard deviation. We want to estimate the variance of that distribution to have an idea of the precision on the computed result. We will compute the variance of the sum or difference and of the product of two independent variables. We will also show how the variance on any function of independent samples can be easily estimated using a first order approximation.

Keep in mind that these derivations only hold for *independent* samples. If some of them are dependent, you should first eliminate them using the equations for the dependencies.

10.4.1 Sum or difference of two independent variables

We have two variables a and b with mean $\overline{a}$ and $\overline{b}$ and variance σ_a^2 and σ_b^2. We compute $a \pm b$ and estimate the corresponding variance $\sigma_{a\pm b}^2$.

$$\begin{aligned}
\sigma_{a\pm b}^2 &= E\left[\Big((a\pm b)-(\overline{a}\pm\overline{b})\Big)^2\right] \\
&= E\left[(a-\overline{a})^2\right] + E\left[(b-\overline{b})^2\right] \pm E\left[2(a-\overline{a})(b-\overline{b})\right] \\
&= E\left[(a-\overline{a})^2\right] + E\left[(b-\overline{b})^2\right] \pm 2E\left[(a-\overline{a})\right]E\left[(b-\overline{b})\right] \\
&= \sigma_a^2 + \sigma_b^2,
\end{aligned} \tag{10.19}$$

because the expectation of $(a-\overline{a})$ is zero. The expectation of the product is the product of the expectations if the variables are independent samples. So in linear combinations the noise adds up, even if the variables are subtracted!

10.4.2 Product of two independent variables

The expectation of a product is the product of the expectations, so we have:

$$\begin{aligned}
\sigma_{ab} &= E\left[\Big(ab-\overline{a}\overline{b}\Big)^2\right] \\
&= E\left[a^2b^2\right] + \overline{a}^2\overline{b}^2 - E\left[2ab\overline{a}\overline{b}\right] \\
&= E\left[a^2\right]E\left[b^2\right] - \overline{a}^2\overline{b}^2
\end{aligned} \tag{10.20}$$

This expression is not very useful, it must be rewritten as a function of $\overline{a}$, $\overline{b}$, σ_a and σ_b. To obtain that, we rewrite a as $\overline{a} + (a - \overline{a})$:

$$\begin{aligned} E\left[a^2\right] &= E\left[(\overline{a} + (a - \overline{a}))^2\right] \\ &= \overline{a}^2 + E\left[(a - \overline{a})^2\right] + E\left[2\overline{a}(a - \overline{a})\right] \\ &= \overline{a}^2 + \sigma_a^2 \end{aligned} \tag{10.21}$$

Substituting this result for $E[a^2]$ and $E[b^2]$ in (10.20) we obtain:

$$\begin{aligned} \sigma_{ab} &= (\overline{a}^2 + \sigma_a^2)(\overline{b}^2 + \sigma_b^2) - \overline{a}^2\overline{b}^2 \\ &= \overline{a}^2\sigma_b^2 + \overline{b}^2\sigma_a^2 + \sigma_a^2\sigma_b^2 \\ &= \overline{a}^2\overline{b^2}\left(\frac{\sigma_a^2}{\overline{a}^2} + \frac{\sigma_b^2}{\overline{b}^2} + \frac{\sigma_a^2\sigma_b^2}{\overline{a}^2\overline{b}^2}\right) \qquad (10.22) \\ &\simeq \overline{a}^2\overline{b^2}\left(\frac{\sigma_a^2}{\overline{a}^2} + \frac{\sigma_b^2}{\overline{b}^2}\right) \qquad (10.23) \end{aligned}$$

The last line is a first order approximation which is acceptable if the relative errors are small. We conclude that when two variables are multiplied the relative variances must be added.

10.4.3 Any function of independent variables

If we can live with first order approximations, the variance of any function of one or more variables can be easily estimated. Consider a function $f(x_1, x_2, \ldots x_n)$ where the x_i are independent samples from distributions characterized by $\overline{x_i}$ and σ_i. Applying first order equations means that f is treated as a linear function:

$$\begin{aligned} E\left[(f(x_1, \ldots x_n) - E\left[f(x_1, \ldots x_n)\right])^2\right] &\simeq E\left[(f(x_1, \ldots x_n) - f(\overline{x_1}, \ldots \overline{x_n}))^2\right] \\ &\simeq E\left[\left(\left.\frac{\partial f}{\partial x_1}\right|_{\overline{x_1}}(x_1 - \overline{x_1}) + \ldots \left.\frac{\partial f}{\partial x_n}\right|_{\overline{x_n}}(x_n - \overline{x_n})\right)^2\right] \\ &= \left(\left.\frac{\partial f}{\partial x_1}\right|_{\overline{x_1}}\right)^2 \sigma_1^2 + \ldots \left(\left.\frac{\partial f}{\partial x_n}\right|_{\overline{x_n}}\right)^2 \sigma_n^2 \end{aligned} \tag{10.24}$$

The first step is a first order approximation: the expectation of a linear function is the function of the expectations. Similarly, the second line is a Taylor expansion, assuming all higher derivatives are zero. The third step is the application of (10.19).

With this approach you can easily verify that the variance on a product or division is obtained by summing the relative variances.

10.5 Expectation of Poisson data contributing to a measurement

Assume the configuration of figure 5.9: two radioactive sources contribute to a single measurement N. We know the a-priori expected values $\bar{a}$ and $\bar{b}$ for each of the sources, and we

know the measured count $N = a + b$. The question is to compute the expected values of a and b given N.

By definition, the expected value equals:

$$E(a|a+b=N) = \frac{\sum_{a=0}^{\infty} p(a|a+b=N)a}{\sum_{a=0}^{\infty} p(a|a+b=N)} \tag{10.25}$$

The denominator should be equal to 1, but if we keep it, we can apply the equation also if p is known except for a constant factor.

The prior expectations for a and b are:

$$p_{\bar{a}}(a) = e^{-\bar{a}}\frac{\bar{a}^a}{a!} \qquad\qquad p_{\bar{b}}(b) = e^{-\bar{b}}\frac{\bar{b}^b}{b!} \tag{10.26}$$

After the measurement, we know that the first detector can only have produced a counts if the other one happened to contribute $N - a$ counts. Dropping some constants this yields:

$$p(a|a+b=N) \sim e^{-\bar{a}}\frac{\bar{a}^a}{a!}\ \ e^{-\bar{b}}\frac{\bar{b}^{N-a}}{(N-a)!} \tag{10.27}$$

$$\sim \frac{\bar{a}^a}{a!}\frac{\bar{b}^{N-a}}{(N-a)!} \tag{10.28}$$

Applying 10.25 yields:

$$E(a|a+b=N) = \frac{\sum_{a=0}^{N} \frac{\bar{a}^a}{a!}\frac{\bar{b}^{N-a}}{(N-a)!}a}{\sum_{a=0}^{N} \frac{\bar{a}^a}{a!}\frac{\bar{b}^{N-a}}{(N-a)!}} \tag{10.29}$$

Let us first look at the denominator:

$$\sum_{a=0}^{N} \frac{\bar{a}^a}{a!}\frac{\bar{b}^{N-a}}{(N-a)!} = \frac{(\bar{a}+\bar{b})^N}{N!} \qquad \text{since } \begin{pmatrix} N \\ a \end{pmatrix} = \frac{N!}{a!(N-a)!} \tag{10.30}$$

A bit more work needs to be done for the numerator:

$$\sum_{a=0}^{N} \frac{\bar{a}^a}{a!}\frac{\bar{b}^{N-a}}{(N-a)!}a = \sum_{a=1}^{N} \frac{\bar{a}^a}{a!}\frac{\bar{b}^{N-a}}{(N-a)!}a \qquad \text{summation can start from 1} \tag{10.31}$$

$$= \sum_{a=1}^{N} \frac{\bar{a}^a}{(a-1)!}\frac{\bar{b}^{N-a}}{(N-a)!} \tag{10.32}$$

$$= \bar{a}\sum_{a=1}^{N} \frac{\bar{a}^{a-1}}{(a-1)!}\frac{\bar{b}^{N-a}}{(N-a)!} \tag{10.33}$$

$$= \bar{a}\sum_{a=0}^{N-1} \frac{\bar{a}^a}{a!}\frac{\bar{b}^{N-1-a}}{(N-1-a)!} \tag{10.34}$$

$$= \bar{a}\frac{(\bar{a}+\bar{b})^{N-1}}{(N-1)!} \tag{10.35}$$

Combining numerator and denominator results in:

$$E(a|a+b=N) = \bar{a}\frac{(\bar{a}+\bar{b})^{N-1}}{(N-1)!}\frac{N!}{(\bar{a}+\bar{b})^N} \tag{10.36}$$

$$= \bar{a}\frac{N}{\bar{a}+\bar{b}} \tag{10.37}$$

10.6 The convergence of the EM algorithm

This section explains why maximizing the likelihood of the complete variables L_x is equivalent to maximizing the likelihood L of the observed variables Q.

First some notations and definitions must be introduced. As in the text, Λ denotes the reconstruction, Q the measured sinogram and X the complete variables. We want to maximize the logarithm of the likelihood given the reconstruction:

$$L(\Lambda) = \ln g(Q|\Lambda) \tag{10.38}$$

The conditional likelihood of the complete variables $f(X|\Lambda)$ can be written as:

$$f(X|\Lambda) = k(X|Q,\Lambda)\; g(Q|\Lambda), \tag{10.39}$$

where $k(X|Q,\Lambda)$ is the conditional likelihood of the complete variables, given the reconstruction and the measurement. These definitions immediately imply that

$$\ln f(X|\Lambda) = L(\Lambda) + \ln k(X|Q,\Lambda). \tag{10.40}$$

The objective function we construct during the E-step is defined as

$$h(\Lambda'|\Lambda) = E\left[\ln f(X|\Lambda')|Q,\Lambda\right], \tag{10.41}$$

which means that we write the log-likelihood of the complete variables as a function of Λ', and that we eliminate the unknown variables X by computing the expectation based on the current reconstruction Λ. Combining (10.40) and (10.41) results in

$$h(\Lambda'|\Lambda) = L(\Lambda') + E\left[\ln k(X|Q,\Lambda')|Q,\Lambda\right] \tag{10.42}$$

Finally, we define a generalized EM (GEM) algorithm. This is a procedure which computes a new reconstruction from the current one. The procedure will be denoted as M. M is a GEM-algorithm if

$$h(M(\Lambda)|\Lambda) \geq h(\Lambda|\Lambda). \tag{10.43}$$

This means that we want M to increase h. We can be more demanding and require that M maximizes h; then we have a regular EM algorithm, such as the MLEM algorithm of section 5.3.2.

Now, from equation (10.42) we can compute what happens with L is we apply a GEM-step to increase the value of h:

$$\begin{aligned} L(M(\Lambda)) - L(\Lambda) \;&=\; h(M(\Lambda)|\Lambda) - h(\Lambda,\Lambda) \\ &\quad + E\left[\ln k(X|Q,\Lambda)|Q,\Lambda\right] - E\left[\ln k(X|Q,M(\Lambda))|Q,\Lambda\right] \end{aligned} \tag{10.44}$$

Because M is a GEM-algorithm, we already know that $h(M(\Lambda)|\Lambda) - h(\Lambda,\Lambda)$ is positive. If we can also show that

$$E\left[\ln k(X|Q,\Lambda)|Q,\Lambda\right] \geq E\left[\ln k(X|Q,M(\Lambda))|Q,\Lambda\right] \tag{10.45}$$

then we have proven that every GEM-step increases the likelihood L. It is possible to prove (10.45) in a general way [4]. Here, we will only verify (10.45) for the particular case of MLEM in emission tomography.

The log-likelihood of the complete variables X, given the measurement Q and the new reconstruction Λ' is given by (appendix 10.5):

$$\ln k(X|Q,\Lambda') = \sum_i \sum_j x_{ij} \ln r_{ij} - r_{ij} \tag{10.46}$$

$$r_{ij} = c_{ij}\lambda'_j \frac{q_i}{\sum_\xi c_{i\xi}\lambda'_\xi} \tag{10.47}$$

The expectation of k given the current reconstruction Λ and the measurement Y is also computed as in appendix 10.5:

$$E\left[\ln k(X|Q,\Lambda')|Q,\Lambda\right] = \sum_i \sum_j n_{ij} \ln r_{ij} - r_{ij} \tag{10.48}$$

$$n_{ij} = c_{ij}\lambda_j \frac{q_i}{\sum_\xi c_{i\xi}\lambda_\xi}. \tag{10.49}$$

The values n_{ij} depend on the current reconstruction, the variables r_{ij} are a function of the new reconstruction Λ'. To see for which r_{ij} the maximum is found we set the derivatives of (10.48) to zero:

$$\frac{\partial}{\partial r_{ij}} \left(\sum_{i'} \sum_{j'} n_{i'j'} \ln r_{i'j'} - r_{i'j'} \right) = 0, \qquad \forall i,j \tag{10.50}$$

$$\Rightarrow \quad \frac{n_{ij}}{r_{ij}} = 1, \qquad \forall i,j \tag{10.51}$$

$$\Rightarrow \quad r_{ij} = n_{ij}, \qquad \forall i,j \tag{10.52}$$

So (10.48) is maximum if $r_{ij} = n_{ij}$. A simple way to achieve that is by setting $\Lambda' = \Lambda$, which confirms that (10.45) holds for emission tomography.

Now we have proved that the GEM-step increases L. It still remains to be shown that the algorithm indeed converges towards the maximum likelihood solution. If you want to know everything about it, you should read the paper by Dempster et al [4].

10.7 The Laplace transform

The Laplace transform is defined as:

$$\mathcal{L}F(t) = f(s) = \int_0^\infty e^{-st} F(t) dt. \tag{10.53}$$

The Laplace transform is very useful in computations involving differential equations, because integrals and derivatives with respect to t are transformed to very simple functions of s. Some of its interesting features are listed below (most are easy to prove). The functions of the time are at the left, the corresponding Laplace transforms are at the right:

$$F(t) \iff f(s) \tag{10.54}$$

$$\frac{dF(t)}{dt} \iff sf(s) - F(0) \tag{10.55}$$

$$e^{at}F(t) \iff f(s-a) \tag{10.56}$$

$$\int_0^t F(u)G(t-u)du \iff f(s)g(s) \tag{10.57}$$

$$\int_0^t F(u)du \iff \frac{f(s)}{s} \tag{10.58}$$

$$1 \iff \frac{1}{s} \tag{10.59}$$

$$e^{at} \iff \frac{1}{s-a} \tag{10.60}$$

Bibliography

[1] JA Sorenson, ME Phelps. "Physics in Nuclear Medicine", WB Saunders Company, Philadelphia, 1987.

[2] MM Ter-Pogossian, "Instrumentation for cardiac positron emission tomography: background and historical perspective", in SR Bergmann and BE Sobel, "Positron emission tomography of the heart", Futura Publishing Company, New York, 1992.

[3] HY Oei, JAK Blokland, Nederlandse Vereniging voor Nucleaire Geneeskunde, "Aanbevelingen Nucleaire Geneeskundige Diagnostiek", Eburon, Delf, ISBN 90-5166-358-7.

[4] AP Dempster, NM Laird, DB Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J R Statist Soc* 1977; 39; 1-38.