ELSEVIER
SAUNDERS

POSITRON
EMISSION
TOMOGRAPHY

PET Clin 2 (2007) 173–190

Advances in PET Image Reconstruction

Andrew J. Reader, PhD[a,*], Habib Zaidi, PhD, PD[b]

PET image reconstruction usually involves the generation of three-dimensional (3D) images of a radiotracer's concentration to estimate physiologic parameters for volumes of interest in vivo. To enhance the functional information, often a time sequence of these 3D images is required so that time-activity curves for each particular volume of interest can be obtained. These time-activity curves can then be fitted with a kinetic model from which functional parameters (such as the metabolic rate of glucose or blood flow) can be estimated. The key point to note is that a PET scanner does not measure the space-time–dependent radiotracer concentration directly. Instead, the annihilation photon pairs arising from the positron emissions are externally detected by the PET scanner and recorded as event histograms (sinograms or projection data) or as a list of recorded photon-pair events (list-mode data). In some cases, the data may even be backprojected to form backprojected images, as a means of data compression, with only minimal loss in spatial precision. Figs. 1 and 2 illustrate the basic formats of data that can be acquired by a PET scanner, indicating how the measured data are indirect measurements of the unknown activity distribution. Indirect measurement is the problem that image reconstruction seeks to solve. Projections/sinograms are the most popular data format, but as more attributes for each PET event are recorded (ie, not only the coordinates of the two detected photons but also energy and timing measurements), list-mode data can become more practical for data storage without loss of information. Although the problem of image reconstruction has conventionally been one of determining the space-time–dependent radioactivity concentration in vivo, there is also a move toward the direct estimation of functional parametric images from the raw PET data.

This work was supported by grant SNSF 3100A0-116547 from the Swiss National Foundation.

[a] School of Chemical Engineering and Analytical Science, The University of Manchester, PO Box 88, Manchester, M60 1QD, UK
[b] Division of Nuclear Medicine, Geneva University Hospital, CH-1211 Geneva, Switzerland
* Corresponding author.
E-mail address: andrew.j.reader@manchester.ac.uk (A.J. Reader).

doi:10.1016/j.cpet.2007.08.001

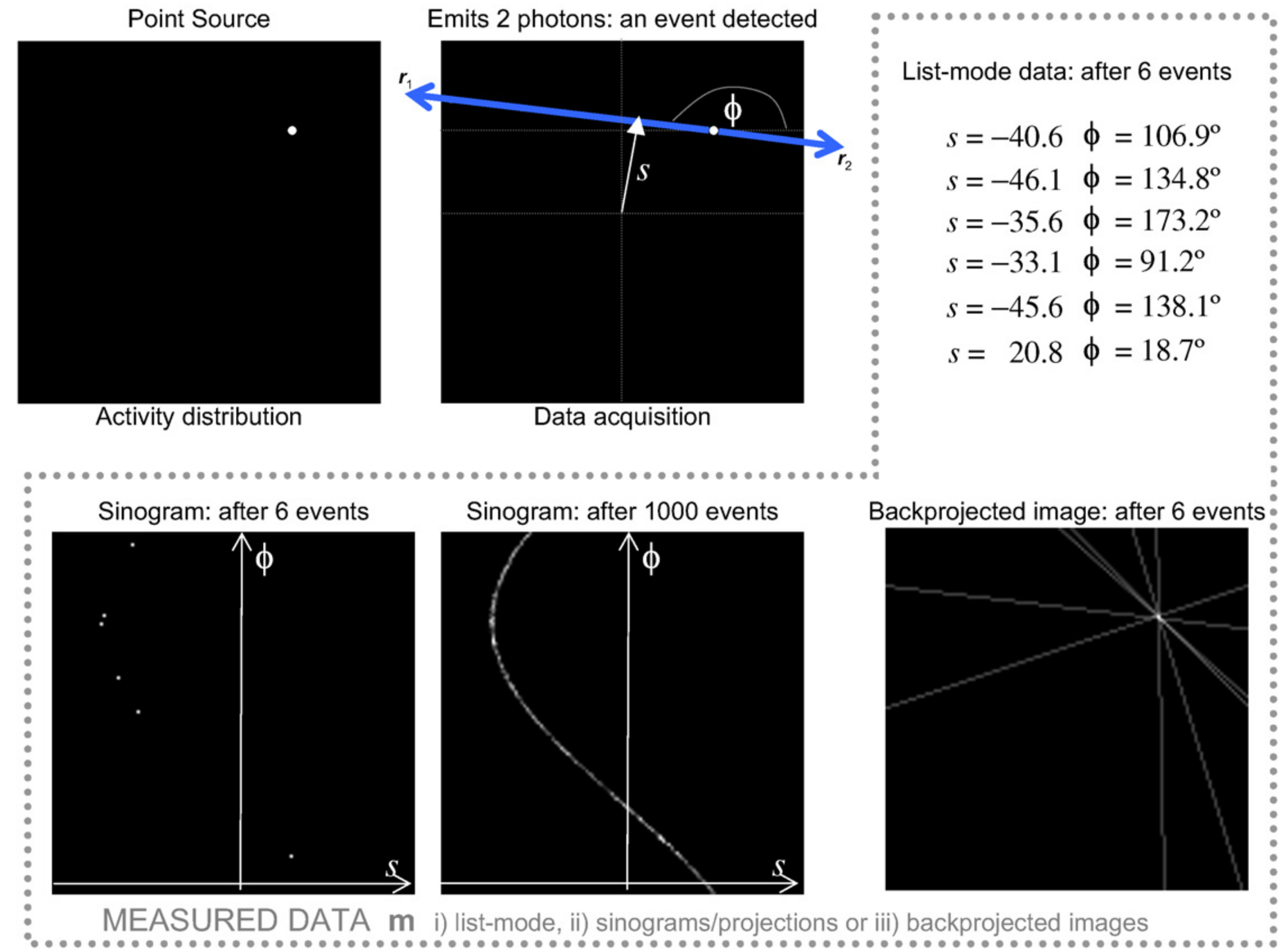

Point Source
Activity distribution
Emits 2 photons: an event detected
r_1
r_2
ϕ
s
Data acquisition
List-mode data: after 6 events
$s = -40.6$ $\phi = 106.9°$
$s = -46.1$ $\phi = 134.8°$
$s = -35.6$ $\phi = 173.2°$
$s = -33.1$ $\phi = 91.2°$
$s = -45.6$ $\phi = 138.1°$
$s = 20.8$ $\phi = 18.7°$
Sinogram: after 6 events
ϕ
s
Sinogram: after 1000 events
ϕ
s
Backprojected image: after 6 events
MEASURED DATA m i) list-mode, ii) sinograms/projections or iii) backprojected images

Fig. 1. The three main formats of PET measured data, shown for the simplified case of 2D PET. (*Top left*) Any radioactivity distribution can be regarded as a collection of point source emitters of different intensities (a single point source example is shown for clarity). (*Top middle*) The radioactivity distribution emits back-to-back photon pairs that may be detected by the PET detectors surrounding the field of view (FOV). Each event is usually associated with a pair of detection points outside the FOV (eg, $\mathbf{r}_1$, $\mathbf{r}_2$), and these two points define a line that can be parameterized by a distance *s* from the origin, and an angle ϕ. (*Top right*) Each detected event can be recorded in a list-mode file (often the raw detection coordinates are recorded, along with any other information such as photon energy). (*Bottom left*) Each detected event can also be accumulated into a sinogram, which has sampling intervals ("bins") in *s* and in ϕ to histogram the events. (*Bottom middle*) Note that after many events, a point source results in a sine wave on the sinogram (hence the name) and that each point (s,ϕ) on the sinogram is approximately equal to a line integral through the radioactivity distribution. (*Bottom right*) An alternative is to backproject the events along their lines of detection. Note that *s* corresponds to the *y′* used in Fig. 4.

Initially this article concerns one-step analytic image reconstruction techniques, which operate on the measured PET projection data to estimate the radiotracer concentration within cubic volume elements (voxels). These methods are applied for each time frame of acquired data, thus providing a time-series of 3D images that can subsequently be used for postreconstruction kinetic analysis to estimate functional parameters. These analytic methods have the advantage of being relatively fast to use and linear and, as a result, they have a good track record for producing reliable quantitative PET images. The article then considers iterative image reconstruction methods, which have the advantages of complete flexibility in modeling the PET data acquisition process and the corresponding freedom in specifying the parameters to estimate. Conventionally, the parameters to estimate are the values (representing radiotracer concentration) in voxels, and the PET data acquisition model is that of taking sets of line-integrals to form projections (ie, the Radon or 3D x-ray transform [1,2]). If voxel values are chosen as the parameters to estimate and if the line-integral imaging model is used, then iterative reconstruction methods (when run to convergence on high-statistics data) often give results broadly comparable to analytic reconstruction. There is an increasing trend, however, toward more fully exploiting the capacities of iterative reconstruction techniques through the use of more accurate and complex modeling of the PET acquisition process. Such modeling can be based on analytic methods or direct measurements (eg, see Refs. [3–7]) or on Monte Carlo methods ([9–11]). In addition, iterative methods easily accommodate different choices of the parameters to be estimated. (Rather than pixel or voxel values, one can consider coefficients for spherically symmetric "blob" basis functions [12,13], "natural pixels" [14–16], or even directly specifying kinetic parameters as the unknowns to be estimated [17,18].)

This article covers the foundations of image reconstruction, with an emphasis on advances and future directions for PET image reconstruction. Recent reviews by Lewitt and Matej [19] and Qi and Leahy [20] and the textbooks by Barrett and Myers [21] and Zaidi [22] also offer excellent comprehensive coverage of image reconstruction algorithms and the foundations of image science appropriate to PET.

Foundations for image reconstruction: representation of the object and the data

Object representation

The unknown space-time radiotracer distribution can be regarded as a continuous function $f(\mathbf{r},\tau)$, specifying the radiotracer concentration (in units such as MBq/mL or mCi/mL) for all spatial locations $\mathbf{r} = [x\,y\,z]^T$ at time τ. It is not practical to measure and reconstruct such a continuous function; this would require essentially infinite levels of sampling and quantities of measured data. Since both sampling and data quantities are limited, a resolution-limited representation of the function f is usually considered, through the use of a set of $j = 1\ldots J$ spatial basis functions $\alpha_j(\mathbf{r})$ and a set of $b = 0\ldots B\text{-}1$ temporal basis functions $\beta_b(\tau)$. In the first instance, consider the time-independent case such that the continuous function f is approximated by

$$f(\mathbf{r}) \approx \sum_{j=1}^{J} c_j \alpha_j(\mathbf{r}) \tag{1}$$

where $\mathrm{c} = \{\mathrm{c_j}\}_J$ is a J-dimensional vector, holding the coefficients for each of the $j = 1\ldots J$ spatial basis functions that cover the field of view (FOV) of the PET scanner. The most common choice for the spatial basis function $\alpha(\mathbf{r})$ is the voxel such that the vector $\mathbf{c}$ simply holds the voxel values and can be interpreted directly (after appropriate rearrangement of the elements) as a two-dimensional (2D) or 3D image. To extend Equation 1 to the time-dependent case, the vector $\mathbf{c}$ can be extended to include the coefficients for the temporal basis functions such that f is now represented by

$$f(\mathbf{r},\tau) \approx \sum_{b=0}^{B-1} \sum_{j=1}^{J} c_{j+bJ} \alpha_j(\mathbf{r}) \beta_b(\tau) \tag{2}$$

where $\mathbf{c}$ is now a JB-dimensional vector, holding the coefficients for each spatial basis function for each of the temporal basis functions. The most common choice for the temporal basis function is the top-hat function such that the function f is composed of distinct time frames, each of which is considered independently from the others. Note that Equation 2 assumes that the function f can be created from factorized spatiotemporal basis functions, rather than more general nonfactorizable spatiotemporal bases $\gamma(\mathbf{r},t)$. For the case of orthogonal basis functions (eg, pixels/voxels or the Fourier basis), the elements of the JB-dimensional vector of coefficients $\mathbf{c}$ are specified by

$$c_{j+bJ} = \int\int f(\mathbf{r},\tau)\alpha_j(\mathbf{r})\beta_b(\tau)d\mathbf{r}d\tau \tag{3}$$

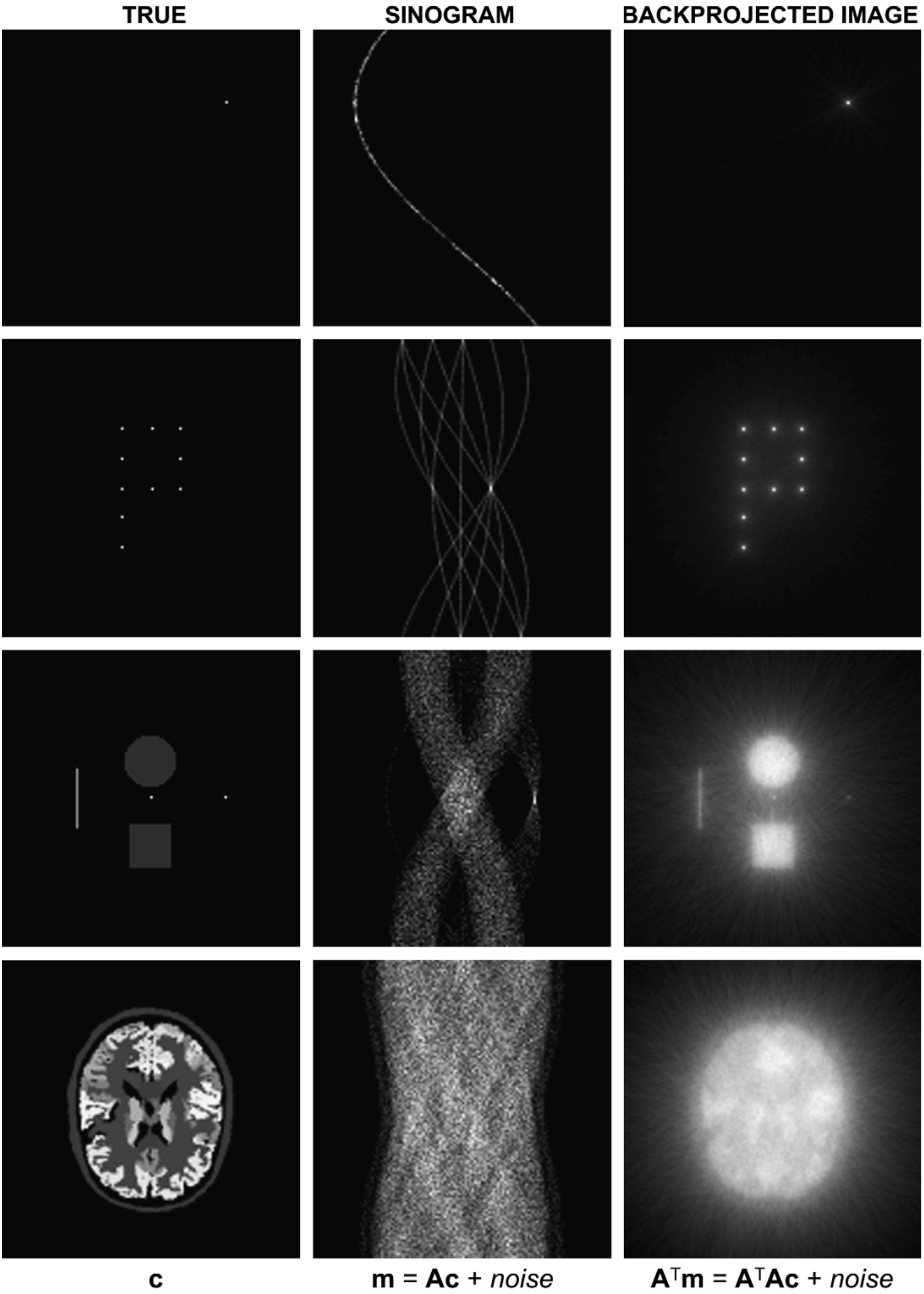
for which the time-independent case (setting $b = 0$) is simply

$$c_j = \int f(\mathbf{r})\alpha_j(\mathbf{r})d\mathbf{r} \tag{4}$$

It should be emphasized that Equations 3 and 4 only hold for orthogonal basis functions (eg, voxels for the spatial functions $\alpha(\mathbf{r})$, top-hat functions for the temporal functions $\beta(\tau)$, or the Fourier basis functions, and so forth). The use of nonorthogonal basis functions is possible, but in such a case, even finding the coefficients vector $\mathbf{c}$ when the function f is actually known is in itself an inverse problem. This, however, reveals yet another advantage of iterative reconstruction methods: nonorthogonal basis functions can be easily used and accounted for by the system model such that the coefficients vector $\mathbf{c}$ is estimated directly. The system model is considered in detail in the following sections.

Data representation

The acquired PET data are usually projections (sinograms or 2D parallel projections) or a list-mode file (although image-based storage such as 3D backprojected images can also be used, see Figs. 1 and 2). List-mode data and projection data can be regarded as a single IT-dimensional vector $\mathbf{m}$, holding the number of counts detected by each of the $i = 1\ldots I$ elements in the measurement space (often these elements are PET detector pairs, commonly referred to as "lines of response" [LORs]) at each of the $t = 1\ldots T$ time sample points. For the case of list-mode data, there is no grouping or resampling ("binning") of the data: information on the exact detector pair i that detected each event, along with the precise time sample point t, is retained by list-mode data (along with additional information such as energy and difference in arrival time of



the photons, in which case the number I of measurement elements would be correspondingly larger). Such precision can be equaled through the use of finely sampled projection data bins (one for each possible event that could be detected), but this often becomes impractical due to the large number of detector pairs employed by state-of-the-art PET systems (eg, $I = 4.5 \times 10^9$ for the high-resolution research tomograph (HRRT) [23]).

Considering first the time-independent case, each element i of the mean $\mathbf{q}$ of the noisy PET measured data $\mathbf{m}$ is modeled by

$$q_i = E\left[m_i\right] = \int f(\mathbf{r})s_i^{\mu}(\mathbf{r})d\mathbf{r} \tag{5}$$

where $s_i{}^{\mu}(\mathbf{r})$ is a function describing the PET scanner sensitivity (for all spatial locations $\mathbf{r}$) to a radiotracer distribution for a given detector pair i, and also for a given linear attenuation coefficient distribution $\mu(\mathbf{r})$. The functions $s_i{}^{\mu}$ $(\mathbf{r})$ can in fact be regarded as probability images, whereby the function value is related to the probability of a positron emission from location $\mathbf{r}$ giving rise to an event being detected by detector pair i (see the later discussion on the system matrix, and Fig. 3). For the time-dependent case, the mean $\mathbf{q}$ of the noisy (ie, count-limited) PET measured data $\mathbf{m}$ is modeled by

$$q_{i+tI} = E\left[m_{I+tl}\right] = \int \left(\int_{\Delta\tau_t} f(\mathbf{r},\tau)d\tau \right) s_i^{\mu}(\mathbf{r})d\mathbf{r} \tag{6}$$

where the τ (time) integral is over the time interval $\Delta\tau_t$ related to the temporal resolution of the time sample t. Conventionally, the set of functions $s_i{}^{\mu}$ ($i = 1\ldots I$) used in Equations 5 and 6 are tubes of response (or LORs) through the FOV, which is why the $i = 1\ldots I$ detector pairs are also referred to as LORs. Using LORs makes Equation 5 equivalent to a line-integral; see Equation 7. In addition, conventionally, the set of spatial basis functions $\alpha_j(\mathbf{r})$ in Equations 1 to 4 are voxels, with the temporal basis functions β_b being top-hat functions. Such a choice for the four-dimensional (4D) volume elements isolates a unique subspace of the space-time volume; however (as is considered later), given the count-limited nature of the acquired data $\mathbf{m}$, there are advantages in considering basis functions that are nonorthogonal and overlap in space-time. Finally, it is worth noting in Equation 5 that if the spatial basis functions $\alpha_j(\mathbf{r})$ correspond to $s_i(\mathbf{r})$ and if these functions are orthogonal, then there is no longer a spatial inverse problem because Equation 4 is immediately obtained, giving all that is necessary to estimate the function f in Equation 1. A similar argument holds for Equations 6 and 3. Hence, it is the difference between the functions $\alpha(\mathbf{r})$ and $s(\mathbf{r})$ (or if they match but are not an orthogonal set of functions) that encapsulates the image reconstruction problem.

Analytic image reconstruction methods

For analytic reconstruction, the approximation of Equation 1 is not assumed; the inverse problem is formulated in a continuous framework and the practical implementation is performed as a discrete approximation of the continuous solution. Also, time-dependent reconstruction is handled as a series of independent "static" reconstructions, so only the time-independent case is considered here. The fundamental assumption of analytic inversion formulae is that a 2D parallel projection representing a set of LORs $p(\mathbf{s}, \hat{\mathbf{u}})$is equal to a set of line integrals through the radioactivity distribution $f(\mathbf{r})$ (the 3D x-ray transform) [1]:

$$p(\mathbf{s},\hat{\mathbf{u}}) = \int_{-\infty}^{+\infty} f(\mathbf{s}+x'\hat{\mathbf{u}})dx' \tag{7}$$

where the 3D vector $\mathbf{r}$ in the imaging volume is decomposed into the 2D parallel projection position vector $\mathbf{s} = [y' \ z']^T$ and the one-dimensional (1D) orientation unit vector $x'\hat{\mathbf{u}}$ (Fig. 4). This line-integral equation can be seen as a special case of Equation 5, whereby the continuous functions $s^{\mu}{}_i(\mathbf{r})$ are now lines through the FOV, with each detector-pair regarded as an LOR i (specified by a displacement $\mathbf{s}$ from the center of the FOV and an orientation $\hat{\mathbf{u}}=(\phi,\theta)$) and with the vector $\mathbf{q}$ replaced by the continuous function p. Effects such as photon attenuation and normalization for detector nonuniformities can be accounted for as multiplicative corrections to these line integrals, and photon scatter and random events need to be subtracted from the measured data before the

Fig. 2. Four examples of activity distributions (*left column*) and their corresponding sinograms (*middle column*) and backprojected images (*right column*) for the simplified case of 2D PET. (*First row*) A positron emitting point source gives rise to a single sine wave of corresponding intensity on the sinogram, and a (1/*r*) distribution (the point response function (PRF)) in the backprojected image. (*Second row*) A collection of point sources gives rise to a corresponding number of sine waves on the sinogram, and a set of PRFs in the backprojected image. (*Third row*) A more general radioactivity distribution can be regarded as a collection of point sources of various intensities; hence, the sinogram is a superposition of sine waves of corresponding intensities for each of these points, and the backprojected image is a superposition of PRFs of corresponding intensities. (*Fourth row*) A brain activity distribution [8]. Since in this case of 2D PET the point response function in the backprojected image is shift-invariant, the simple model of convolution can be used and reconstruction can be achieved by deconvolution. Annotations using the vector **c** and the matrix **A** are added for reference in later sections of the text.

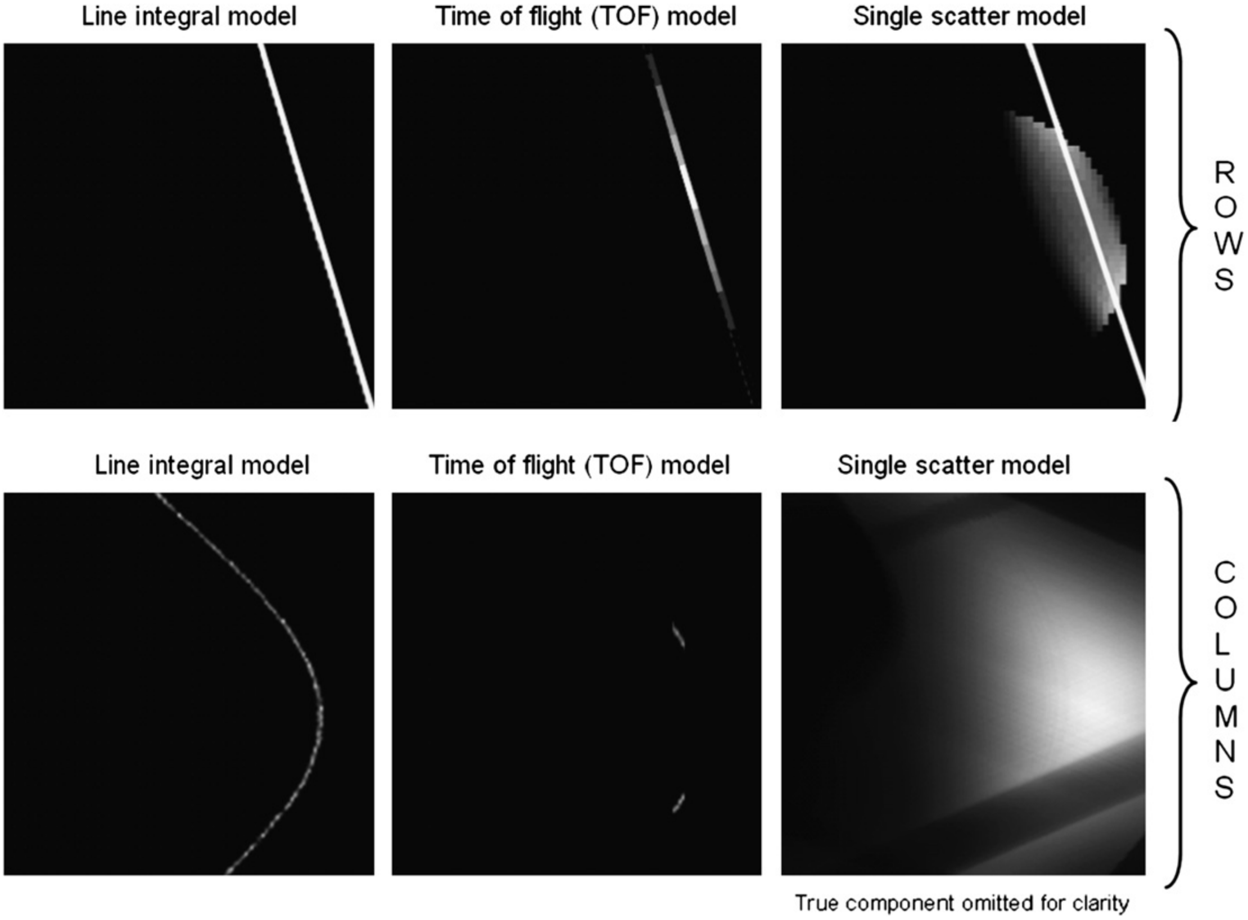


Fig. 3. Illustrative example of rows and columns of the system matrix **A** for different imaging models (the simple case of using pixels with 2D PET is considered). (*Top row*) Three examples of a single row *i* of the matrix **A** (a row is equivalent to an image) are shown for the case of the line integral model, a model using time-of-flight information, and incorporating scatter into the system matrix. Forward modeling of a given radioactivity distribution (represented by **c**) to obtain an element q_i of the mean data **q** merely involves taking a pixel-by-pixel product of one of these row images with the **c** image and summing all values (ie, performing a scalar product). (*Bottom row*) Three examples of a single column *j* of the system matrix (in forward modeling, to get mean data **q** from **c**, the columns are weighted by each point source intensity [ie, each pixel value in the **c** image] and summed to produce the overall expected data **q**). A simple and approximate way of including an image-space resolution model would be to convolve the row images, or for a simple detector-space resolution model, one could convolve the column data. Note that the image values are not to scale and are for illustrative purposes only.

approximation of Equation 7 can be considered as appropriate. Hence, the measured data $\mathbf{m} = \{m_i \ldots m_I\}$ are approximately regarded as proportional to Equation 7 according to

$$KA_iN_i(m_i - \rho_i - \sigma_i) \approx \int_{-\infty}^{+\infty} f(\mathbf{s} + x'\hat{\mathbf{u}})dx' \qquad (8)$$

where A_i is the attenuation correction factor for LOR i (calculated by taking the exponential of a 3D x-ray transform of the attenuation map or from the ratio of blank to transmission scans); N_i is the normalization correction factor for LOR i; and ρ_i and σ_i are estimates of the randoms and scatter on the projections. The global scale factor K accounts for corrections such as radioactive decay and detector dead time. Equation 8 assumes that the positron-electron annihilation event occurred exactly along the LOR that joins the two detectors in detector pair i, thus ignoring photon acollinearity, positron range, and parallax error.

The inversion of Equation 7 (ie, to find f, given p) can be found through the central section theorem, which relates the projection data p to the image f [24]: a central plane of the 3D Fourier transform $F(\mathbf{k})$ of the 3D true activity $f(\mathbf{r})$ is equal to the 2D Fourier transform $P(\mathbf{k})$ of the 2D parallel projection data at the same orientation $\hat{\mathbf{u}} = (\phi, \theta)$. Considering the case of 2D reconstruction, the central section theorem corresponds to the 1D Fourier transform of a 1D parallel projection being equal to a single line through the 2D Fourier transform of f. It becomes apparent that the 2D Fourier transform of f can be constructed by the superposition of these 1D Fourier transforms (one from each projection angle), but that this superposition of the many lines at many

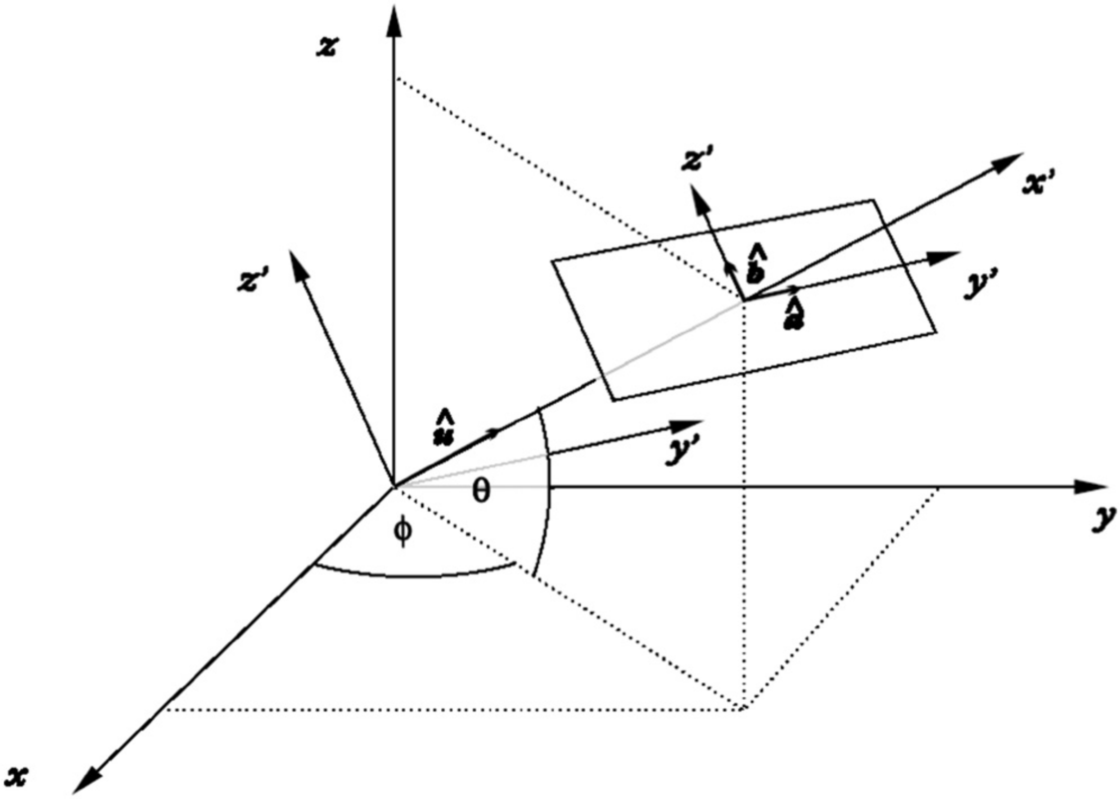


Fig. 4. A 2D parallel projection (2D PP), at orientation $\hat{\mathbf{u}}=(\phi,\theta)$ where ϕ is the azimuthal angle and θ is the copolar angle. The y'–z' plane ($\mathbf{s} = [y'\ z']^T$) holds the LORs, each characterized by a distance **s** from the center of the FOV, and by an orientation $\hat{\mathbf{u}}=(\phi,\theta)$, along which the 3D radioactivity distribution $f(\mathbf{r})$ is considered to be integrated. For 3D PET, the direct (θ=0) 2D PPs are complete and sufficient to reconstruct $f(\mathbf{r})$, but the oblique (|θ|>0) 2D PPs are truncated, due to the limited axial extent of 2D PET scanners. One of the sinograms from Fig. 2 would be obtained by considering the case of θ =0 and z' =0, leaving just y' and ϕ as the two coordinates for a single, direct sinogram.

angles will give rise to a $1/|r|$ density of contributions to the 2D Fourier transform. The ramp filter is used (equal simply to $|r|$) to compensate for the varying contributions in the frequency domain. Rather than performing a direct reconstruction by way of the Fourier domain, it is often more straightforward to filter the projections and then backproject them.

Various commonly used analytic reconstruction methods are described in the literature [25,26], including the simple backprojection method (which results in a blurred version of the object distribution, as was shown in Fig. 2) [27]; backprojection followed by filtering (BPF) [28]; and filtered backprojection (FBP)/convolution backprojection. The BPF method usually relies on angle-limited backprojection of the list-mode or projection data to produce a shift-invariant point response function (PRF) that can then be deconvolved by Fourier methods.

In the particular case of 3D PET, because there is sufficient information to reconstruct the 3D activity distribution using only the direct projections (ie, those resulting from LORs within the same detector ring: θ = 0), the oblique projections (|θ|>0) are redundant and are not required to reconstruct the activity distribution. Given the count-limited nature of the acquired data, however, using these oblique projections has the potential to effectively increase the sensitivity of the scanner and thus reduce the noise in the reconstructed image. When the oblique projections are not truncated, they can be reconstructed by FBP using the Colsher filter [29]. This algorithm can be implemented as FBP or BPF (the backprojection operation is 3D in both cases). However, for real PET systems that have a limited axial extent, the 3D reconstruction problem is complicated by the fact that all the oblique (|θ|>0) 2D projections are truncated (ie, they are incompletely measured). One way to handle this is by recognizing that the direct (θ = 0) projections allow complete reconstruction of the activity distribution. The activity distribution reconstructed from these direct projections can then be reprojected to obtain the missing portions of the oblique projections. In practice, as much as 40% of the total reconstruction time can be spent in estimating and then backprojecting the unmeasured projection data. The algorithm combining this forward projection step with the Colsher filter is a 3D FBP algorithm referred to as the 3D reprojection (3D RP) method [30].

An alternative exact approach to solve the problem of truncated 2D projections is the Fast Volume Reconstruction (FAVOR) algorithm [31]. FAVOR, however, presents different noise propagation properties compared with other approaches (ie, it does not uniformly weight the data, unlike the Colsher filter). With FAVOR, simple ramp filtering can be used but at the expense of suboptimally weighting the data, attributing a much greater weight to the direct projections than to the oblique ones.

Although 3D RP works well and has served as a "gold standard" for 3D analytic reconstruction algorithms, it requires significant computational resources (compared with, for example, a series of 2D FBP slice reconstructions). As a result, a number of approximate methods have been proposed. Most of the well-known approaches involve rebinning the data from the oblique projections into 2D direct sinogram data sets, thus allowing the use of 2D reconstruction techniques and resulting in a significant decrease in computation time. These algorithms rebin the 3D data into a stack of ordinary 2D direct sinograms, and the image is reconstructed in 2D slice by slice, for instance using the standard 2D FBP or iterative algorithms. The method used to rebin the data from the oblique sinograms into the smaller 2D data sets is the crucial step in any rebinning algorithm and has a significant influence on the resolution and noise in the reconstructed image.

The simplest rebinning algorithm is single-slice rebinning [32]. It is based on the simple approximation that an oblique LOR can be projected to the direct plane halfway between the two end points of the LOR. Hence, single-slice rebinning assigns counts from an oblique LOR (in which the detectors are on different axial PET detector rings) to the sinogram of the transaxial slice lying midway axially between the two rings. This approximation is acceptable only near the central axis of the scanner and for small apertures. For larger apertures and for greater displacements from the axis of the scanner, single-slice rebinning results in position-dependent axial blurring.

Several alternative methods have been suggested to provide more accurate rebinning. Lewitt and colleagues [33] proposed a more refined rebinning approach, the multislice rebinning algorithm, that is more accurate than single-slice rebinning but suffers from instabilities in the presence of noisy data. The Fourier rebinning (FORE) algorithm is a more sophisticated algorithm in which oblique rays are binned to a transaxial slice using the frequency-distance relationship [34] of the data in Fourier space [35]. Although rebinning can be done exactly (using, for example, the FOREX or FOREPROJ methods) [36], a considerable speed advantage is gained through using the approximate FORE approach. Another exact rebinning approach referred to as FORE-J was proposed and is based on the fact that the 3D x-ray transform of a function is a solution to a second-order partial differential equation (John's equation) [37]. FORE-J is easy to implement, given that it has the same structure as FORE with a small modification, although it involves adding a small correction to each oblique projection before rebinning. Matej and Kazantsev [38] revived the interest for the direct inversion of the 3D x-ray transform by proposing sophisticated interpolation techniques in the Fourier domain, provided that the oblique projection data are not truncated. More recently, the approach derived by Defrise was extended to derive a 3D exact rebinning formula in Fourier space, leading to an iterative reprojection algorithm (iterative FOREPROJ) that allows estimation of unmeasured oblique projections using the whole set of measured projections [39].

Iterative image reconstruction methods

The primary limitation of the analytic inversion methods just described is their reliance on the line-integral Equation 7 (ie, the 3D x-ray transform), rather than the more general Equation 5. Iterative methods not only accommodate more complex (and hence realistic) models of the acquired PET data but also easily allow the use of nonorthogonal basis functions.

The system matrix and the parameters

Two of the fundamental components to an iterative reconstruction algorithm are (1) the definition of the parameters to estimate (usually a set of values that provides a representation of the radiotracer concentration, such as the vector **c** in Equation 1, and (2) the definition of the system model **A** (which describes the relation between the radiotracer distribution and the mean of the measured data). Note that it is the mean of the measured data that is normally modeled, and with iterative algorithms, there is substantial flexibility in how this mean is modeled and how it is parameterized. Because the system model (mapping from the parameters to the mean) is usually time invariant, this case is considered here. The elements of the system matrix (or the system model) $\mathbf{A} = \{a_{ij}\}_{I\times J}$ are defined by

$$a_{ij} = \int s_i^{\mu}(\mathbf{r})\alpha_j(\mathbf{r})d\mathbf{r} \tag{9}$$

Considering the previous Equations 1 and 5 along with Equation 9, the matrix-vector relation

$$\mathbf{q} = \mathbf{A}\mathbf{c} \tag{10}$$

is obtained, which allows a mean data set **q** to be "predicted" for any given set of parameters **c** (where a set of parameters **c** represents a radioactivity distribution). A mean data set **q** can never be observed in practice but can be regarded as the average number of counts that would theoretically be obtained in each detector pair i by repeating precisely the same PET scan (of a radioactivity concentration specified by **c**) an infinite number of times.

It is perhaps helpful to visualize the contents of the matrix **A** for a case in which the vector **c** contains the coefficients of voxels (ie, voxel values such that **c** can be simply rearranged to form a 2D or 3D image). In such a case, the rows of the matrix **A** can be regarded as images, with values proportional to the probability of a positron emission in each of the $j = 1\ldots J$ voxels giving rise to an event being detected by detector pair (or LOR) i (see Fig. 3). Hence, for the case of the x-ray transform (line integrals), the probability image is simply an image of a single line passing through the FOV (the line could be found by any of the many ray-tracing methods; eg, see Refs. [40–42]). If positron range were to be included, then the line would need to be broader; for example, by convolving the ray-traced line with a positron-range kernel if the approximation of shift-invariant blurring is made (there are fast methods for achieving such approaches; eg, see Ref. [43]). If time-of-flight information were to be included, then a Gaussian function (with mean equal to the estimated annihilation location and with variance linked to the timing resolution [44]) can be used to modulate the values along the line. In general, for the most accurate system modeling, essentially every element of these probability images that make up the rows of **A** would be nonzero. (If, for example, scatter were to be included in the system model [7], then there is still a finite probability that positron emissions from voxels remote from the LOR i will result in photons ultimately detected along i.)

In a similar manner, the columns of the matrix **A** can be visualized as "data images" that show the entire system response (ie, the mean data **q**) to a single point source emitter. Thus, for the case of the 3D x-ray transform, each column of **A** would be a set of sinograms corresponding to the mean set of 3D sinograms that would be obtained from a point source acquisition. Consequently, there is considerable impetus to design system matrices that more accurately model the detection process within **A**, through Monte Carlo simulation [45] or even from actual point source measurements [5]. (The recent high-definition PET (HD-PET) [46] relies on this kind of improved system modeling to deliver marked improvements in image quality.)

Objective functions

It is the vector **c** (which represents the radioactivity concentration, Equation 1) that needs to be estimated from the noisy measured PET data **m**. An iterative image reconstruction algorithm aims to find a vector $\mathbf{c}^k$, which produces a vector $\mathbf{q}^k$ (by way of the system matrix **A**) that matches in some way with **m** (where the superscript k on **c** and **q** denotes an estimate number, usually corresponding to an iteration number). Most algorithms seek to minimize a distance measure between the measured data **m** and the vector $\mathbf{q}^k$, which can be specified as a function of $\mathbf{c}^k$. This function is called the objective function. The two most widely used objective functions in PET are least squares (LS) and maximum likelihood (ML). The LS objective concerns finding the vector $\mathbf{c}^k$, which minimizes the following function:

$$O_{LS}\left(\mathbf{c}^k\right) = \sum_{i=1}^{I}\left(m_i - q_i^k\left(\mathbf{c}^k\right)\right)^2 \tag{11}$$

where, as previously specified in Equation 10,

$$q_i^k\left(\mathbf{c}^k\right) = \sum_{j=1}^{J} a_{ij}c_j^k \tag{12}$$

The ML objective is based on a statistical model of the noisy data vector **m**, and is therefore often preferred in PET. The probability of obtaining a measurement m_i (ie, a number of counts m in detector pair i) if the mean value was q_i^k (ie, if the vector of parameters was $\mathbf{c}^k$) is given by the Poisson distribution:

$$\Pr\left(m_i\middle|q_i^k\right) = \frac{\left(q_i^k\right)^{m_i}}{m_i!}\exp\left[-q_i^k\right] \tag{13}$$

The value of q_i^k that maximizes this probability is simply $q_i^k = m_i$. For a given $\mathbf{c}^k$, a whole set of q_i^k values is obtained (ie, the entire vector $\mathbf{q}^k$, by way of Equation 10 or 12). The likelihood of obtaining the vector **m** if the mean was $\mathbf{q}^k$ is defined by the product of these Poisson probabilities.

$$\Pr\left(\mathbf{m}\middle|\mathbf{q}^k\right) = \prod_{i=1}^{I}\frac{\left(q_i^k\right)^{m_i}}{m_i!}\exp\left[-q_i^k\right] \tag{14}$$

In order, to maximize Equation 14, each of the terms in the product must individually be maximized. Therefore, the concept of the ML estimate becomes evident: it corresponds to the choice of $\mathbf{c}^k$, which gives a vector $\mathbf{q}^k$ by way of Equation 10, which gives a maximum value for the likelihood

Equation 14. Note that if the measured data had been direct measurements of the parameters of interest (ie, if **A** was the unit matrix **I**), then the trivial result of $\mathbf{c}^k = \mathbf{q}^k = \mathbf{m}$ is obtained as the ML estimate. It is only the introduction of a nontrivial **A** that creates the reconstruction problem. The logarithm of Equation 14 is usually taken to simplify the mathematics (but taking the logarithm will not alter the location of the maximum of the function):

$$\ln\Pr\left(\mathbf{m}\middle|\mathbf{q}^k\right) = \sum_{i=1}^{I} m_i \ln q_i^k - \sum_{i=1}^{I} q_i^k - \sum_{i=1}^{I} \ln m_i! \qquad (15)$$

Equation 15 is called the log likelihood. Maximizing the log likelihood also corresponds to minimizing the Kullback-Leibler distance measure between the noisy vector **m** and the estimate of the mean $\mathbf{q}^k$. Many other objective functions exist [20], but the main two encountered in PET are the ML and the LS objectives as mentioned earlier. It is also worth noting that if a Gaussian model of the measured data statistics is used, then the ML objective corresponds to the LS objective.

Direct algorithms

Before exploring further the iterative methods that achieve minimization or maximization of a particular objective function, it is worth noting the existence of direct methods for estimating the coefficients vector **c**. To clarify the discussion, **c** will be regarded as holding the values for pixel or voxel basis functions such that any estimate $\hat{\mathbf{c}}$ can be interpreted directly as a 2D or 3D image. In essence, a solution vector $\mathbf{c}^s$ to the set of linear equations

$$\mathbf{m} = \mathbf{A}\mathbf{c}^s \qquad (16)$$

is sought, although for many reasons, no such solution vector $\mathbf{c}^s$ exists (eg, **A** is rarely square and **m** is noisy). Singular value decomposition of **A** into the three matrices $\mathbf{U\Sigma V}^T$, however, can allow a pseudoinverse for a direct LS estimate of **c**:

$$\hat{\mathbf{c}} = \mathbf{V}\Sigma^{-1}\mathbf{U}^T\mathbf{m} \qquad (17)$$

where **U** and **V** are orthogonal matrices and **Σ** is a diagonal matrix (holding the singular value spectrum). This approach has been successfully demonstrated in 2D PET [47]; however, for 3D PET, the size of **A** is typically of the order of $10^9 \times 10^7$ elements, which makes such an approach computationally impractical at the present time [48] (but worthy of consideration in the future). Alternatively, applying the transpose of matrix **A** to both sides of Equation 16 results in

$$\mathbf{A}^T\mathbf{m} = \mathbf{A}^T\mathbf{A}\mathbf{c}^s \qquad (18)$$

where the overall matrix $\mathbf{A}^T\mathbf{A}$ is square, with columns corresponding to the PRF for each possible point source location, and $\mathbf{A}^T\mathbf{m}$ can be interpreted as a backprojected image. Consequently, the estimate

$$\hat{\mathbf{c}} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{m} \qquad (19)$$

can be found, which corresponds to the familiar normal equations for LS fitting. The inverse of $\mathbf{A}^T\mathbf{A}$ could be found by eigenvector decomposition:

$$\left(\mathbf{A}^T\mathbf{A}\right)^{-1} = \left(\mathbf{EDE}^T\right)^{-1} = \mathbf{ED}^{-1}\mathbf{E}^T \qquad (20)$$

where **E** is the orthogonal matrix of eigenvectors and **D** is a diagonal matrix holding the eigenvalues. As with singular value decomposition, the problem is that $\mathbf{A}^T\mathbf{A}$ is a huge, nonsparse matrix (typically $>10^{14}$ elements), which does not allow straightforward computation of the eigenvectors. If, however, the case of an angular constrained backprojection along LORs is considered (as in the BPF method; see the Analytic Image Reconstruction Methods section), then $\mathbf{A}^T\mathbf{A}$ corresponds to a convolution (ie, its columns contain shifted copies of the PRF or kernel) and, hence, the matrix of eigenvectors contains the Fourier basis functions, and the diagonal matrix **D** simply holds the modulation transfer function (the Fourier transform of the PRF/kernel) along its diagonal. This means that an LS solution can be directly found by Fourier methods:

$$\hat{\mathbf{c}} = \mathbf{ED}^{-1}\mathbf{E}^T\mathbf{A}^T\mathbf{m} \qquad (21)$$

Equation 21 corresponds to (1) backprojection of the measured data ($\mathbf{A}^T\mathbf{m}$); (2) Fourier transforming the image (applying the transpose of the eigenvector matrix **E**); (3) multiplying point by point in the Fourier domain by the inverse of the transfer function ($\mathbf{D}^{-1}$); and finally, (4) inverse Fourier transforming (applying **E**). Noise can be controlled by modulating, truncating, or adding an offset to the Fourier domain filter contained in the diagonal of $\mathbf{D}^{-1}$.

To summarize, **A** and $\mathbf{A}^T\mathbf{A}$ are huge matrices that do not readily lend themselves to pseudoinversion methods, and Fourier domain methods can be applied only when the columns of $\mathbf{A}^T\mathbf{A}$ correspond to shifted copies of the same PRF (ie, a convolution). Therefore, although such direct approaches would be desirable (to avoid issues of convergence), they pose a significant computational challenge. Iterative techniques, on the other hand, often only need row or column access of **A** and often avoid explicit storage of the matrix **A** through on-the-fly calculation, through factorization of **A** into manageable components (see later discussion), or both.

Iterative algorithms

Most iterative reconstruction algorithms in PET rely on finding the gradient of the objective function, which is a basic principle of optimization. For example, if the objective function were quadratic, then the gradient would be linear. Any estimate of the location of the maximum (or minimum) of the objective function could then be improved by simply adding the value of the gradient at the current location. A key question is how much of the gradient to add on (known as the step size). Hence, the basic form of an iterative reconstruction algorithm is often

$$c_j^{k+1} = c_j^k + \lambda^k \frac{\partial O(\mathbf{c}^k)}{\partial c_j^k} \tag{22}$$

where λ^k is the step size. For the case of c corresponding to voxel values, the gradient can be regarded as an image. For the log-likelihood objective, the gradient image is given by differentiating Equation 15:

$$\frac{\partial}{\partial c_j^k} O(\mathbf{c}^k) = \sum_{i=1}^{I} a_{ij} \frac{m_i}{q_i^k} - \sum_{i=1}^{I} a_{ij} \tag{23}$$

Equation 23 is just a backprojection of what can be considered as correction factors for each i (the $m_i/q_i{}^k$ ratio, which indicates whether the current estimate $\mathbf{q}^k$ of the mean of the data is above or below the noisy data $\mathbf{m}$) minus the sensitivity image $(\sum_{i=1}^{I} a_{ij})$. The sensitivity image corresponds to the backprojection of every possible system LOR; that is, the summation of all the "images" that make up all the rows of the matrix **A**. Referring again to Fig. 4, for the line-integral model, it would mean adding together every possible line through the FOV. There is effectively a choice in how much of this gradient image to add to obtain the next estimate of **c**. If a step size of

$$\lambda^k = \frac{c_j^k}{\sum_{i=1}^{I} a_{ij}} \tag{24}$$

is chosen and substituted into Equation 22 along with Equation 23 to obtain

$$c_j^{k+1} = c_j^k + \frac{c_j^k}{\sum_{i=1}^{I} a_{ij}} \left(\sum_{i=1}^{I} a_{ij} \frac{m_i}{q_i^k} - \sum_{i=1}^{I} a_{ij} \right) \tag{25}$$

then the well known expectation maximization (EM) algorithm is obtained [49,50]:

$$c_j^{k+1} = \frac{c_j^k}{\sum_{i=1}^{I} a_{ij}} \sum_{i=1}^{I} a_{ij} \frac{m_i}{q_i^k} \tag{26}$$

In matrix-vector form, this can be written as

$$\mathbf{c}^{k+1} = \frac{\mathbf{c}^k}{\mathbf{A}^T \mathbf{1}} \mathbf{A}^T \frac{\mathbf{m}}{\mathbf{q}^k} \tag{27}$$

where **1** is a vector with all elements set to one, and element-by-element vector multiplication and division is assumed for simplicity of notation, but where matrix-vector multiplication remains conventional (eg, see Ref. [51]). The notation of Equation 27 becomes useful when data corrections and decomposition of the system matrix are used (see the Decomposition of the System Matrix, Including Attenuation, Normalization, Scatter, and Randoms section) to prevent excess writing of summations and subscripts. Hence, the core iterative process of the EM algorithm of Equation 26 involves (1) forward projection of the current estimate ($\mathbf{q}^k = \mathbf{A}\mathbf{c}^k$) to model the mean; (2) taking the ratio of the measured data to this mean ($\mathbf{m}/\mathbf{q}^k$); (3) backprojecting this ratio [$\mathbf{A}^T(\mathbf{m}/\mathbf{q}^k)$] to form a multiplicative correction image; and (4) multiplying the correction image by the current estimate of $\mathbf{c}^k$ to obtain the new estimate $\mathbf{c}^{k+1}$. It is also necessary to normalize for the number of contributions to the correction image, hence the division by $\mathbf{A}^T\mathbf{1}$, which represents the backprojection of every possible LOR (ie, the superposition of every row of the system matrix **A**, as previously mentioned, to obtain the sensitivity image).

Variants on Equation 26 include over-relaxation [52–54] and methods that perform a line search (a 1D optimization of λ to find the step size that maximizes the log likelihood). This latter approach corresponds to the steepest ascent algorithm [55]. Another important modification is to divide the data vector **m** into L subsets $S_1 \ldots S_L$, whereby from the L subsets of the vector **m**, L updates of the estimate of **c** can be performed:

$$c_j^{k,l+1} = \frac{c_j^{k,l}}{\sum_{i \in S_l} a_{ij}} \sum_{i \in S_l} a_{ij} \frac{m_i}{q_i^{k,l}} \tag{28}$$

This modification leads to a much more computationally efficient update because the amount of processing per update is approximately a factor L smaller. This block-iterative version of the EM algorithm is referred to as "ordered subsets" EM (OSEM) [56]. The problem with OSEM is that convergence to an ML estimate is in general lost, and instead, a limit cycle corresponding to the number of subsets is encountered. This loss of convergence can be overcome through consideration of previous corrective images in each update [57] or through the use of a relaxation parameter Λ^k as is done in the row-action ML algorithm (RAMLA) [58,59]:

$$c_j^{k,l+1} = c_j^{k,l} + \Lambda^k \frac{c_j^{k,l}}{\sum\limits_{i\in S_l} a_{ij}} \left(\sum_{i\in S_l} a_{ij} \frac{m_i}{q_i^{k,l}} - \sum_{i\in S_l} a_{ij} \right) \tag{29}$$

This algorithm is simply Equation 25 but with the use of subsets and relaxation. Another interesting example is to combine over-relaxation with subsets (such as the high–over-relaxation single-projection method [60]).

Turning now to the gradient of the LS objective,

$$\frac{\partial}{\partial c_j^k} O_{LS}\left(\mathbf{c}^k\right) = 2\sum_{i=1}^{I} a_{ij}\left(m_i - q_i^k\left(\mathbf{c}^k\right)\right) \tag{30}$$

it can be seen that algorithms such as the algebraic reconstruction technique [61] use a form of this gradient:

$$c_j^{k,l+1} = c_j^{k,l} + \lambda^k a_{ij} \frac{\left(m_i - q_i^{k,l}\right)}{\sum\limits_j \left(a_{ij}\right)^2} \tag{31}$$

This algorithm is equivalent to taking each LOR *i* as a subset of the vector **m** for each update and using the gradient of the LS objective with a relaxed step size.

Faster gradient methods include the preconditioned conjugate gradient method, which usually takes the form of the following steps for PET [20,55,62,63]. First, a modified gradient image is obtained by multiplication of the current gradient image vector $\mathbf{g}^k$ by a preconditioning matrix **C**:

$$\mathbf{d}^k = \mathbf{C}^k \mathbf{g}^k \tag{32}$$

where the preconditioning matrix is normally taken to be diagonal, with elements inspired by the step size of the gradient used in the EM algorithm (the step size given in Equation 24), such as

$$\mathbf{C}^k = diag\left\{ \left(c_j^k + \delta\right) \Big/ \sum_i a_{ij} \right\} \tag{33}$$

where δ is a small constant. Another step size is then found through

$$\beta^{k-1} = \frac{\left(\mathbf{g}^k - \mathbf{g}^{k-1}\right)^T \mathbf{d}^k}{\mathbf{g}^{k-1^T} \mathbf{d}^{k-1}} \tag{34}$$

which is used for the updating step of a vector **a** defined by

$$\mathbf{a}^k = \mathbf{d}^k + \beta^{k-1}\mathbf{a}^{k-1} \tag{35}$$

which is finally used to update the vector of parameters to be estimated:

$$\mathbf{c}^{k+1} = \mathbf{c}^k + \alpha^k \mathbf{a}^k \tag{36}$$

Equation 36 then requires a 1D optimization to find the α^k that maximizes the objective function; this is nontrivial for the ML case and, often, a Newton-Raphson method (with a positivity constraint) is employed. More recently, even faster gradient techniques using subsets have been devised [64].

Updating methods

It is unfortunate that the choice of algorithm is often dictated by the type of data to reconstruct from: list-mode data or projection data. For list-mode data, a so-called RAMLA (such as EM and its variants OSEM and row-action ML algorithm) is often essential for practical reasons. A row-action method is one that accesses the rows of the system matrix one at a time, which corresponds directly to accessing individual events in the list-mode data. Furthermore, for list-mode data, it is impractical to consider algorithms that require every possible LOR to be accessed for each update [65]. As a result, methods such as ART (Equation 31) are impractical for list-mode data because the zero values in the vector **m** need to be accounted for in each update. Likewise, column-based (ie, voxel-based) methods are impractical for list-mode data because the entirety of the randomly ordered list-mode file needs to be considered for each voxel update. For projection data, the choice of a column-based algorithm, whereby a whole column of the system matrix is required for a single update of a voxel, is practical. Examples of column-based algorithms include space-alternating generalized EM [66], iterative coordinate ascent, grouped-coordinate ascent, and successive over-relaxation [67] and related methods. As stated, these approaches can be readily implemented for projection data but tend to be impractical for list-mode data.

Regularization

A key problem often encountered in iterative reconstruction is the ill conditioning of the inverse problem such that a trivial perturbation in the measured data **m** gives rise to a nontrivial perturbation in the image estimate. This issue can be understood through consideration of the singular value spectrum of the matrix **A** (or in simpler terms, the modulation transfer function for the diagonalization of $\mathbf{A}^T\mathbf{A}$ in the very special case in which this is a convolution matrix). The values in the spectrum decay, indicating the decreasing efficiency with which the various singular vector components of the image are measured by the PET scanner. Any algorithm that progresses toward effectively achieving any type of inversion of **A** will be performing the equivalent of inverting the decaying spectrum of singular values (or inverting the transfer function in the special case of Fourier), and division by the increasingly small values in the singular value spectrum can cause significant noise amplification. Hence, regularization methods such as postsmoothing, early termination of the iterative process, or bayesian priors (used to modify the ML objective to a maximum a posteriori [MAP] objective) are often used to counteract this. Of these possibilities, early termination of the iterative process is by far the most popular choice, but this can lead to problems (if quantification is essential) because spatially variant and object-dependent convergence occurs with nonlinear updating algorithms such as EM. MAP methods or postsmoothing (after reaching convergence) can counteract these problems. Qi and Leahy [20] provide a good review of MAP regularization. A significant amount of prior information on image smoothness, however, can be accounted for in the modeling of the mean of the data—that is, a choice of spatial basis function that is no longer a voxel but rather a smooth function (such as spherically symmetric basis functions; for example, "blobs" [12,68], or even a cluster of voxels [69]).

Decomposition of the system matrix, including attenuation, normalization, scatter, and randoms

As previously mentioned, the system matrix **A** for 3D PET can be prohibitively large. To make the matrix manageable, on-the-fly calculation can be performed (ie, the system matrix elements are calculated as and when they are required) or symmetry and compression techniques can be exploited (eg, see Refs. [70,71]). A popular approach is to decompose the matrix **A** into components that are individually easy to calculate or easy to store (eg, see Refs. [72,73]). An example of this approach would be

$$\mathbf{A} = \mathbf{NDLXH} \tag{37}$$

where **X** is a geometric projection matrix (eg, this could perform line-integrals such that each row of **X** can be interpreted as an image of a line through the FOV); where **H** models image space resolution (eg, positron range); **D** models resolution effects in the detector space; **L** is a diagonal matrix accounting for attenuation along each LOR; and **N** is a diagonal matrix giving the inverse of the normalization correction factor for each LOR. Specific examples include using a convolution operation for **H** (whereby the columns of the matrix contain shifted copies of the resolution kernel), a ray tracing algorithm (such as the Siddon method [41] or that of Joseph [40]) for **X**, and a convolution matrix for **D**. Note that the matrix **X** can be modified to account for time-of-flight information, through replacement of the implicit uniform weighting along each line by a Gaussian probability density, with a mean and variance corresponding to the photon pair detection time difference and the timing resolution (see Fig. 3). The advantage of the decomposition of the form shown in Equation 37 is that each matrix can be stored (in the case of the diagonal matrices **N** and **L** and possibly **X** if symmetry is exploited) or easily calculated on the fly (in the case of **D**, **X**, and **H**). Inclusion of scatter and randoms is normally done in an additive way such that the mean of the data is modeled by

$$\mathbf{q} = \mathbf{NDLXHc}+\mathbf{s}+\mathbf{r} \tag{38}$$

where **s** and **r** are estimates of the mean of the scatter and randoms, respectively, in the measurement space. Scatter could be included in the matrix [7,71], which is strictly the correct way to handle scatter [74,75] (because **s** is normally not known without a first estimate of the true activity distribution). Using the model of Equation 38, the EM algorithm of Equation 27 becomes

$$\mathbf{c}^{k+1} = \frac{\mathbf{c}^{k}}{\mathbf{H}^T\mathbf{X}^T\mathbf{L}^T\mathbf{D}^T\mathbf{N}^T\mathbf{1}}\mathbf{H}^T\mathbf{X}^T\mathbf{L}^T\mathbf{D}^T\mathbf{N}^T\frac{\mathbf{m}}{\mathbf{NDLXHc}^{k}+\mathbf{s}+\mathbf{r}} \tag{39}$$

Equation 39 handles factors such as attenuation and scatter by means of modeling (rather than by explicit corrections such as subtraction of the scatter) and, as such, the statistical form of the data vector **m** is preserved as Poisson. Hence, an algorithm such as Equation 39 is often referred to as an "ordinary Poisson" version of the EM algorithm. There are important advantages to such an approach; for example, the attenuation is included as a forward modeling step rather than as a precorrection (as was done for Equation 8), which more optimally weights the acquired data (hence the name "attenuation weighted" OSEM [76]).

Patient motion can also be included through a system modeling approach (eg, using the Polaris system to supply a feed of movement information synchronized with the acquired PET data) [77].

Four-dimensional methods

4D image reconstruction methods aim to more optimally use the acquired PET data to reconstruct a time series of images or even to directly estimate the kinetic/functional parameters of interest. There are two basic approaches: (1) estimation of coefficients for temporally extensive basis functions to include the time dimension in the reconstruction and (2) estimation of the time-dependent functional parameters (eg, blood flow) directly from the data. The motivation behind both approaches is to avoid the independent time-frame reconstructions used for conventional dynamic PET, which are characterized by relatively poor signal-to-noise ratio arising from the very limited number of counts available for each time frame.

Reconstruction of independent time frames ignores temporal correlations, whereas the use of temporally extensive basis functions allows each time point in the time series reconstruction to draw from more, if not all, of the acquired data. Examples of these types of method include Nichols [78], who used spline temporal basis functions; Matthews [79], who used singular value decomposition to derive a set of basis functions; Reader [69], who estimated the temporal basis functions from the data during reconstruction; and Verhaeghe [80], who considered five-dimensional basis functions (the further dimension coming from gating the data). The principle of such methods when applied to time-dependent PET data **m** can be represented by

$$\mathbf{c}^{k+1} = \frac{\mathbf{c}^{k}}{\mathbf{B}^{T}\mathbf{A}_{t}^{T}\mathbf{1}}\mathbf{B}^{T}\mathbf{A}_{t}^{T}\frac{\mathbf{m}}{\mathbf{A}_{t}\mathbf{B}\mathbf{c}^{k}+\mathbf{b}} \qquad (40)$$

where the system matrix $\mathbf{A}_t$, often assumed to be time independent, holds $t = 1 \ldots T$ copies of the time-independent matrix **A** previously considered (see The System Matrix and the Parameters section); the matrix **B** holds the temporal basis functions; and **b** holds the time-dependent estimates of the scatter and randoms. An example result from this kind of algorithm is shown in Fig. 5.

There is also an increasing move toward task-oriented image reconstruction such that the final purpose of the images is accounted for within the reconstruction process (eg, to deliver higher resolution and better quality images suited to the task (Fig. 6) or to directly estimate the kinetic parameters of interest). One specific example is for the case of [^{18}F]fluorodeoxyglucose (FDG), whereby images of the metabolic rate of glucose are sought, rather than images of the time course of FDG concentration in tissue. The work of Kamasak [17] is a good example of this, which estimated the kinetic parameters for the two-tissue compartment model at each voxel by using a parametric iterative coordinate descent algorithm (each voxel's parameter set is updated to maximize the likelihood). Such approaches are based on the original idea of Carson (the EM parametric image reconstruction algorithm) [81]; methods that extend this approach further (eg, to include estimation of the input function) have followed [82]. Another recent approach is that of Fu and colleagues [18], whereby the parameters of a Patlak plot [83] (a plot normally performed as a postreconstruction analysis to determine the irreversible radiotracer uptake) are estimated directly within a preconditioned conjugate gradient reconstruction.

Summary

Image reconstruction for PET has progressed from the use of FBP algorithms based on analytic inversion formulae (reliant on a simple line-integral model) to the use of iterative algorithms that allow more accurate modeling of the PET data (in terms of the acquisition process and the statistical noise). Iterative methods (which are often terminated before convergence and often use a positivity constraint) usually deliver visually improved image quality that perhaps cannot be matched by direct methods, even when they use the same system model (eg, pseudoinversion by singular value

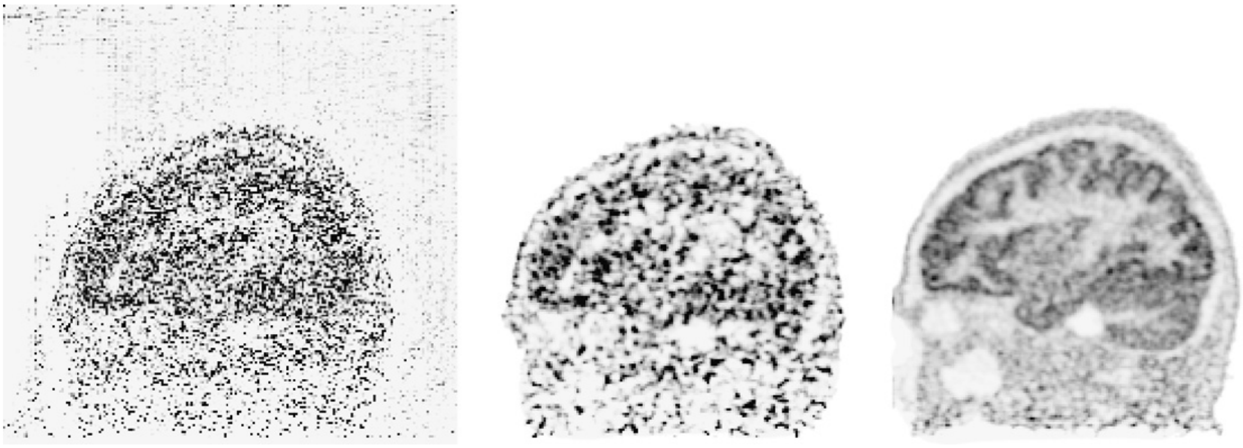


Fig. 5. Impact of 4D reconstruction methodology for one 5-minute frame of a 60-minute [^{11}C]flumazenil study using the HRRT. (*Left*) Sagittal slice, reconstructed using an EM algorithm based only on the line-integral model (ie, A = X). (*Middle*) The same slice reconstructed with list-mode EM including a resolution model with the line-integral model (ie, A = XH). (*Right*) The same slice reconstructed using a 4D method that is able to benefit from all 60 minutes of data without compromising temporal resolution (although it is dependent on the type of temporal basis functions chosen or estimated). (The raw HRRT PET data are *courtesy of* CEA–Service Hospitalier Frédéric Joliot, Michel Bottlaender, Orsay, France.)

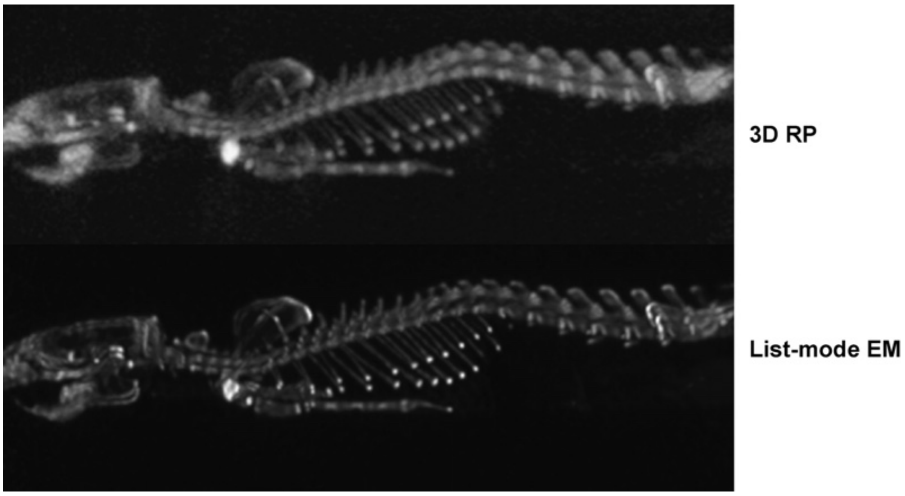


Fig. 6. Comparison of maximum-intensity projections of an F 18 skeletal rat scan generated using analytic (3D reprojection) reconstruction and iterative list-mode EM reconstruction. The iterative EM method benefits from improved modeling of the acquired PET data, which significantly improves image quality.

decomposition). It is well known, however, that nonlinear algorithms such as EM demonstrate spatially variant and object-dependent convergence properties, and any reconstruction based on early termination of the iterative process must be used with caution. The final purpose of the images needs to be carefully considered when selecting a reconstruction algorithm and its associated parameters (iterations, regularization, and so forth). If the purpose of the PET images is to determine whether tumors are present (lesion detection), then incompletely converged EM images may serve well in this role. On the other hand, if quantitative measures of blood flow or receptor density were needed, then a converged reconstruction (or even an FBP reconstruction) would be more appropriate.

There is still further scope for progress in PET image reconstruction, which can perhaps be summarized by the following three points:

1. Hardware developments mean that the acquired PET data vector **m** is increasingly rich in information and can now include high-precision measurements of photon detection positions, timing, photon arrival time differences (time of flight), and photon energy. In addition, measurements of patient movement and respiratory and cardiac gating information are increasingly available as parallel streams of information [77]. To use these data in their richest form would suggest the need for list-mode data storage and reconstruction, rather than binning of data into projections (which can sometimes impose a level of sampling that compromises the original measurement information).
2. Improved accuracy in the definition of the probabilistic system transfer matrix **A** is needed. This may involve accurate Monte Carlo modeling of the positron emission, photon emission, transport and detection processes [45], with the modeling making use of CT or MR images for anatomic information. The photon interactions need to be considered not only in the individual patient but also for the detection hardware. Ideally, there needs to be a comprehensive probabilistic mapping relating the radioactivity distribution to each possible list-mode event.
3. More careful definition of the parameters to be estimated is needed. Are images of radiotracer concentration required or should images of blood flow, receptor density, or metabolic rates be directly estimated from the raw PET data? There is no reason why PET reconstruction cannot be tailored to address specific clinical research questions through direct estimation of the pertinent parameters of interest.

References

[1] Solman DC. The x-ray transform. J Math Anal Appl 1976;56:61–83.

[2] Defrise M. Solution to the three-dimensional image reconstruction problem from two-dimensional parallel projections. J Opt Soc Am A 1993;10:869–77.

[3] Mumcuoglu EU, Leahy RM, Cherry SR, et al. Accurate geometric and physical response modelling for statistical image reconstruction in high resolution PET. Presented at the IEEE Nuclear Science Symposium and Medical Imaging Conference. Anaheim, November 3–9, 1996. Conference Record. 1996;3:1569–73.

[4] Selivanov VV, Picard Y, Cadorette J, et al. Detector response models for statistical iterative image reconstruction in high resolution PET. IEEE Trans Nucl Sci 2000;47:1168–75.

[5] Panin VY, Kehren F, Michel C, et al. Fully 3-D PET reconstruction with system matrix derived from point source measurements. IEEE Trans Med Imaging 2006;25:907–21.

[6] Alessio AM, Kinahan PE, Lewellen TK. Modeling and incorporation of system response functions in 3-D whole body PET. IEEE Trans Med Imaging 2006;25:828–37.

[7] Markiewicz PJ, Tamal M, Julyan PJ, et al. High accuracy multiple scatter modelling for 3D whole body PET. Phys Med Biol 2007;52:829–47.

[8] Reilhac A, Evans AC, Gimenez G, et al. Creation and application of a simulated database of dynamic [18F]MPPF PET acquisitions incorporating inter-individual anatomical and biological variability. IEEE Trans Medical Imaging 2006; 25:1431–9.

[9] Beekman FJ, de Jong HW, van Geloven S. Efficient fully 3-D iterative SPECT reconstruction with Monte Carlo-based scatter compensation. IEEE Trans Med Imag 2002;21:867–77.

[10] Rafecas M, Mosler B, Dietz M, et al. Use of a Monte Carlo-based probability matrix for 3-D iterative reconstruction of MADPET-II data. IEEE Trans Nucl Sci 2004;51:2597–605.

[11] Lazaro D, El Bitar Z, Breton V, et al. Fully 3D Monte Carlo reconstruction in SPECT: a feasibility study. Phys Med Biol 2005;16:3739–54.

[12] Lewitt RM. Alternatives to voxels for image representation in iterative reconstruction algorithms. Phys Med Biol 1992;37:705–16.

[13] Matej S, Lewitt RM. Efficient 3D grids for image reconstruction using spherically-symmetric volume elements. IEEE Trans Nucl Sci 1995;42:1361–70.

[14] Buonocore MH, Brody WR, Macovski A. A natural pixel decomposition for two-dimensional image reconstruction. IEEE Trans Biomed Eng 1981;28:69–78.

[15] Baker JR, Budinger TF, Huesman RH. Generalized approach to inverse problems in tomography: image reconstruction for spatially variant systems using natural pixels. Crit Rev Biomed Eng 1992;20:47–71.

[16] Vandenberghe S, Staelens S, Byrne CL, et al. Reconstruction of 2D PET data with Monte Carlo generated system matrix for generalized natural pixels. Phys Med Biol 2006;51:3105–25.

[17] Kamasak ME, Bouman CA, Morris ED, et al. Direct reconstruction of kinetic parameter images from dynamic PET data. IEEE Trans Med Imaging 2005;24:636–50.

[18] Fu L, Wang G, Qi J. Direct maximum a posteriori reconstruction of Patlak parametric image for fully 3D dynamic PET. Proceedings of the 9th international meeting on fully 3-D image reconstruction in radiology and nuclear medicine. Lindau (Germany), July 9–13, 2007. Kachelriess M, Beekman F, editors. 2007. p. 197–200.

[19] Lewitt RM, Matej S. Overview of methods for image reconstruction from projections in emission computed tomography. Proc IEEE Inst Electr Electron Eng 2003;91:1588–611.

[20] Qi J, Leahy RM. Iterative reconstruction techniques in emission computed tomography. Phys Med Biol 2006;51:R541–78.

[21] Barrett HH, Myers K. Foundations of image science. Hoboken (NJ): John Wiley & Sons; 2003.

[22] Zaidi H. Quantitative analysis in nuclear medicine imaging. New York: Springer; 2006. p. 564.

[23] Wienhard K, Schmand M, Casey ME, et al. The ECAT HRRT: performance and first clinical application of the new high resolution research tomograph. IEEE Trans Nucl Sci 2002;49:104–10.

[24] Natterer F. The mathematics of computerized tomography. New York: Wiley; 1986.

[25] Budinger TF, Gullberg GT. Three-dimensional reconstruction in nuclear medicine emission imaging. IEEE Trans Nucl Sci 1974;NS-21:2–20.

[26] Kak AC, Slaney M. Principles of computerized tomographic imaging. New York: IEEE Press; 1988.

[27] Kuhl KE, Edwards RQ. Image separation radioisotope scanning. Radiology 1963;80:653–61.

[28] Chu G, Tam K-C. Three-dimensional imaging in the positron camera using Fourier techniques. Phys Med Biol 1977;22:245–65.

[29] Colsher JG. Fully three-dimensional positron emission tomography. Phys Med Biol 1980;25:103–15.

[30] Kinahan PE, Rogers JG. Analytic 3D image reconstruction using all detected events. IEEE Trans Nucl Sci 1989;36:964–8.

[31] Defrise M, Townsend DW, Clack R. FAVOR: a fast reconstruction algorithm for volume imaging in PET. Presented at the IEEE Nuclear Science Symposium and Medical Imaging Conference. Santa Fe, November 2–9, 1991. Conference Record. 1992:1919–23.

[32] Daube-Witherspoon ME, Muehllehner G. Treatment of axial data in three-dimensional PET. J Nucl Med 1987;28:1717–24.

[33] Lewitt RM, Muehllehner G, Karp JS. Three-dimensional image reconstruction for PET by multi-slice rebinning and axial image filtering. Phys Med Biol 1994;39:321–39.

[34] Edholm PR, Lewitt RM, Lindholm B. Novel properties of the Fourier decomposition of the sinogram. Proc Soc Photo Opt Instrum Eng 1986;671:8–18.

[35] Defrise M, Kinahan PE, Townsend DW, et al. Exact and approximate rebinning algorithms for 3-D PET data. IEEE Trans Med Imaging 1997;16:145–58.

[36] Liu X, Defrise M, Michel C, et al. Exact rebinning methods for three-dimensional PET. IEEE Trans Med Imaging 1999;18:657–64.

[37] Defrise M, Liu X. A fast rebinning algorithm for 3D positron emission tomography using John's equation. Inverse Probl 1999;15:1047–65.

[38] Matej S, Kazantsev IG. Fourier-based reconstruction for fully 3-D PET: optimization of interpolation parameters. IEEE Trans Med Imaging 2006; 25:845–54.

[39] Ben Bouallegue F, Crouzet F, Comtat C, et al. Exact and approximate Fourier rebinning algorithms for the solution of the data truncation problem in 3-D PET. IEEE Trans Med Imaging 2007;26:1001–9.

[40] Joseph PM. An improved algorithm for reprojecting rays through pixel images. IEEE Trans Med Imaging 1982;1:192–6.

[41] Siddon R. Fast calculation of the exact radiological path for 3-D CT. Med Phys 1985;12:252–5.

[42] Zhao H, Reader AJ. Fast ray-tracing technique to calculate line integral paths in voxel arrays. Presented at the IEEE Nuclear Science Symposium and Medical Imaging Conference. Seattle, October 19–25, 2003. Conference Record. 2003;4:2808–12.

[43] Reader AJ, Ally S, Bakatselos F, et al. One-pass list-mode EM algorithm for high-resolution 3-D PET image reconstruction into large arrays. IEEE Trans Nuc Sci 2002;49:693–9.

[44] Moses WW. Time of flight in PET revisited. IEEE Trans Nucl Sci 2003;50:1325–30.

[45] Zaidi H. Relevance of accurate Monte Carlo modeling in nuclear medical imaging. Med Phys 1999;26:574–608.

[46] Casey M. Point spread function reconstruction in PET. Knoxville (TN): White paper, Siemens Molecular Imaging USA, Inc.; 2007.

[47] Selivanov VV, Lecomte R. Fast PET image reconstruction based on SVD decomposition of the system matrix. IEEE Trans Nuc Sci 2001;48: 761–7.

[48] Selivanov VV, Lepage MD, Lecomte R. List-mode image reconstruction for real-time PET imaging. J Vis Comm Imag Repres 2006;17:630–46.

[49] Llacer J. Tomographic image reconstruction by eigenvector decomposition: its limitations and areas of applicability. IEEE Trans Med Imaging 1982;1:34–42.

[50] Shepp LA, Vardi Y. Maximum likelihood reconstruction for emission tomography. IEEE Trans Med Imaging 1982;1:113–22.

[51] Barrett H, Wilson D, Tsui B. Noise properties of the EM algorithm: I. Theory. Phys Med Biol 1994;39:833–46.

[52] Lewitt RM, Muehllehner G. Accelerated iterative reconstruction for positron emission tomography based on the EM algorithm for maximum likelihood estimation. IEEE Trans Med Imaging 1986;5:16–22.

[53] Fessler JA. Penalized weighted least-squares image reconstruction for positron emission tomography. IEEE Trans Med Imaging 1994;13: 290–300.

[54] Hwang D, Zeng GL. Convergence study of an accelerated ML-EM algorithm using bigger step size. Phys Med Biol 2006;51:237–52.

[55] Kaufman L. Maximum likelihood, least squares, and penalized least squares for PET. IEEE Trans Med Imaging 1993;12:200–14.

[56] Hudson HM, Larkin RS. Accelerated image reconstruction using ordered subsets of projection data. IEEE Trans Med Imaging 1994;13: 601–9.

[57] Hsiao IT, Rangarajan A, Khurd P, et al. An accelerated convergent ordered subsets algorithm for emission tomography. Phys Med Biol 2004;49: 2145–56.

[58] Browne J, de Pierro AB. A row-action alternative to the EM algorithm for maximizing likelihood in emission tomography. IEEE Trans Med Imaging 1996;15:687–99.

[59] de Pierro AR, Yamagishi MEB. Fast EM-like methods for maximum “a posteriori” estimates in emission tomography. IEEE Trans Med Imaging 2001;20:280–8.

[60] Schmidlin P, Bellemann ME, Brix G. Iterative reconstruction of PET images using a high-overrelaxation single-projection algorithm. Phys Med Biol 1997;42:569–82.

[61] Herman GT, Lent A, Rowland SW. ART: mathematics and applications. A report on the mathematical foundations and on the applicability to real data of the algebraic reconstruction techniques. J Theor Biol 1973;42:1–32.

[62] Kaufman L, Neumaier A. PET regularization by envelope guided conjugate gradients. IEEE Trans Med Imaging 1996;15:385–9.

[63] Chinn G, Huang SC. A general class of preconditioners for statistical iterative reconstruction of emission computed tomography. IEEE Trans Med Imaging 1997;16:1–10.

[64] Li Q, Asma E, Ahn S, et al. A fast fully 4-D incremental gradient reconstruction algorithm for list mode PET data. IEEE Trans Med Imaging 2007; 26:58–67.

[65] Kadrmas DJ. LOR-OSEM: statistical PET reconstruction from raw line-of-response histograms. Phys Med Biol 2004;49:4731–44.

[66] Fessler JA, Hero AO III. Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. IEEE Trans Image Proc 1995;4:1417–29.

[67] Sauer K, Bouman C. A local update strategy for iterative reconstruction from projections. IEEE Trans Sign Proc 1993;41:534–48.

[68] Matej S, Lewitt RM. Practical considerations for 3-D image reconstruction using spherically symmetric volume elements. IEEE Trans Med Imaging 1996;15:68–78.

[69] Reader AJ, Sureau FC, Comtat C, et al. Joint estimation of dynamic PET images and temporal basis functions using fully 4D ML-EM. Phys Med Biol 2006;51:5455–74.

[70] Hong IK, Chung ST, Kim HK, et al. Ultra fast symmetry and SIMD-based projection-backprojection (SSP) algorithm for 3-D PET image reconstruction. IEEE Trans Med Imaging 2007;26: 789–803.

[71] Rehfeld N, Alber M. A parallelizable compression scheme for Monte Carlo scatter system matrices in PET image reconstruction. Phys Med Biol 2007;52:3421–37.

[72] Mumcuoglu EU, Leahy RM, Cherry SR. Bayesian reconstruction of PET images: methodology and performance analysis. Phys Med Biol 1996;41: 1777–807.

[73] Qi J, Leahy RM, Cherry SR, et al. High-resolution 3D Bayesian image reconstruction using the microPET small-animal scanner. Phys Med Biol 1998;43:1001–13.

[74] Zaidi H, Koral KF. Scatter modelling and compensation in emission tomography. Eur J Nucl Med Mol Imaging 2004;31:761–82.

[75] Tamal M, Reader AJ, Markiewicz PJ, et al. Noise properties of four strategies for incorporation of scatter and attenuation information in PET reconstruction using the EM-ML algorithm. IEEE Trans Nuc Sci 2006;53:2778–86.

[76] Michel C, Liu X, Sanabria S, et al. Weighted schemes applied to 3D-OSEM reconstruction in PET. Presented at the IEEE Nuclear Science Symposium and Medical Imaging Conference. Seattle, October 24–30, 1999. Conference Record. 1999;3:1152–7.

[77] Rahmim A, Rousset OG, Zaidi H. Strategies for motion tracking and correction in PET. PET Clinics, in press.

[78] Nichols TE, Qi J, Asma E, et al. Spatiotemporal reconstruction of list-mode PET data. IEEE Trans Med Imaging 2002;21:396–404.

[79] Matthews J, Bailey D, Price P, et al. The direct calculation of parametric images from dynamic PET data using maximum-likelihood iterative reconstruction. Phys Med Biol 1997;42:1155–73.

[80] Verhaeghe J, D'Asseler Y, Staelens S, et al. Reconstruction for gated dynamic cardiac PET imaging using a tensor product spline basis. IEEE Trans Nuc Sci 2007;54:80–91.

[81] Carson RE, Lange K. The EM parametric image reconstruction algorithm. J Am Stat Assoc 1985; 80:20–2.

[82] Yetik IS, Qi J. Direct estimation of kinetic parameters from the sinogram with an unknown blood function. Presented at the 3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro. Arlington, VA, April 6–9, 2006. 2006: 295–8.

[83] Patlak CS, Blasberg RG, Fenstermacher JD. Graphical evaluation of blood-to-brain transfer constants from multiple-time uptake data. J Cereb Blood Flow Metab 1983;3:1–7.