Journal of Cerebral Blood Flow & Metabolism
23:490–498 © 2003 The International Society for Cerebral Blood Flow and Metabolism
Published by Lippincott Williams & Wilkins, Inc., Philadelphia

On the Undecidability Among Kinetic Models: From Model Selection to Model Averaging

Federico E. Turkheimer, Rainer Hinz, and Vincent J. Cunningham

Imaging Research Solutions Ltd., Cyclotron Building, Hammersmith Hospital, London, United Kingdom

Summary: This article deals with the problem of model selection for the mathematical description of tracer kinetics in nuclear medicine. It stems from the consideration of some specific data sets where different models have similar performances. In these situations, it is shown that considerate averaging of a parameter's estimates over the entire model set is better than obtaining the estimates from one model only. Furthermore, it is also shown that the procedure of averaging over a small number of "good" models reduces the "generalization error," the error introduced when the model selected over a particular data set is applied to different conditions, such as subject populations with altered physiologic parameters, modified acquisition protocols, and different signal-to-noise ratios. The method of averaging over the entire model set uses Akaike coefficients as measures of an individual model's likelihood. To facilitate the understanding of these statistical tools, the authors provide an introduction to model selection criteria and a short technical treatment of Akaike's information–theoretic approach. The new method is illustrated and epitomized by a case example on the modeling of [^{11}C]flumazenil kinetics in the brain, containing both real and simulated data. **Key Words:** PET—Kinetic modeling—Model selection—Model averaging—Akaike information coefficient—Akaike weights.

Hirotugu Akaike introduced the Akaike information coefficient (AIC) in 1973 (Akaike, 1973). Since then, the AIC has become a standard tool for model selection in regression analysis.

One interesting and usually overlooked feature of AIC is that the coefficient is an estimator of the Kullback-Leibler (K-L) distances of the models from the true generating mechanism of the data (Kullback and Leibler, 1951), regardless of whether the true model is in the set. This property is indeed quite powerful and has important consequences. First, it neatly formalizes the vague notion that no "true" model for the data exists. Truly the very word *model* implies simplification and idealization. Modeling has been defined as the process of constructing idealized representations that capture important aspects of complex biologic and physical systems (Chatfield, 1995). Second, if the "true" model does not exist, then the notion of a model selection procedure that selects the "best" model and discards others may lose appeal. An alternative possibility is that there may be more than one model that can be regarded as useful (i.e., able to provide a sufficiently close approximation to the data for the purpose at hand) (Chatfield, 1995). Third, recent work in statistics has shown that procedures that select the best model from a set, particularly in regression analysis, are intrinsically unstable (Breiman, 1996). This means that the model selection procedure introduces an amount of variance comparable to the error due to parameter estimation. Regression procedures can be "stabilized," however, by appropriate averaging of parameter values over the entire model set (Breiman, 1996).

These considerations can be developed, with minimal conceptual and technical effort, into a new paradigm of "model weighted averaging" where estimates of the parameter of interest are obtained from the entire model class or a suitable subset (Buckland et al., 1997). This article introduces a model averaging approach into the context of classical compartmental modeling. Its evaluation will be carried out using illustrative data generated with positron emission tomography (PET).

Received July 19, 2002; final version received November 7, 2002; accepted November 13, 2002.

Rainer Hinz was supported by a research grant from the Deutsche Forschungsgemeinschaft (Geschäftszeichen HI 769/1–1).

Address correspondence and reprint requests to Federico E. Turkheimer, Neuropathology Department, Faculty of Medicine, Imperial College of Science, Technology and Medicine Charing Cross Campus, Fulham Palace Road, London, W6 8RF, UK; e-mail: federico.turkheimer@ic.ac.uk

Abbreviations used: AIC, Akaike information coefficient; BIC, Bayesian information coefficient; K-L, Kullback-Leibler; ML, maximum likelihood; MSE, mean squared error; PET, positron emission tomography; TAC, time–activity curve.

DOI: 10.1097/01.WCB.0000050065.57184.BB

THEORETICAL BACKGROUND

The problem of estimating the dimension of a model occurs in various forms in applied statistics (e.g., multiple linear and nonlinear regression, factor analysis, time-series modeling). Most of the foundations of the subject were laid down in the mid 1970s. The C_p criterion of Mallows (1973) and the PRESS criterion of Allen (1974) were introduced for use with linear regression. At the same time more widely applicable approaches were developed, including AIC, the cross-validation method (Stone, 1974), the Bayesian information coefficient (BIC) (Schwarz, 1978), and minimum description length (Rissanen, 1978). Work has proceeded on the lines of these methods and modified as well as new criteria have appeared in the literature (Burnham and Anderson, 1998).

Buckland et al. (1997) advocated the approach of model averaging introduced here, but the approach had been considered in the original work of Akaike (1978*b*, 1979). The model averaging approach is based on AIC by reasons of its information-theoretic background, its formal links to maximum likelihood (ML) theory, and its fundamental connections to the area of the foundations of statistics. These aspects will be briefly reviewed in the following sections. Since a model-averaging paradigm exists in the Bayesian framework, the relation between AIC, BIC, and Bayesian inference will be also sketched for the sake of completeness and clarity.

The Kullback-Leibler distance

The K-L distance between the models f and g is defined as:

$$I(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x|\theta)}\right) dx \qquad (1)$$

The function f is considered fixed, whereas g varies over a space of models indexed by its parameter vector θ. X is the sample space where $x \in X$. $I(f,g)$ is amenable to two interpretations. One is of "information loss" when g is used to approximate f, and the second is of "distance" of g from the data-generating process f.

$I(f,g)$ is a fundamental measure of information and is related to other information quantifiers (Soofi, 1994). If f is defined on the same parameter space of g (i.e., $f(x) = g(x|\theta_0)$), then for $\theta \to \theta_0$ $I(f,g)$ is equal to the Fisher information matrix (Kullback and Leibler, 1951). $I(f,g)$ has relations to the general concept of entropy $H(x) = -\log(f(x))$ originated by Boltzmann (1877), to Shannon's entropy $H(x) = \int f(x) \log (f(x))\, dx$ in communication theory (Shannon, 1948) and can be interpreted as the cross-entropy between f and g (Soofi, 1994).

Estimation of Kullback-Leibler information

Akaike (1973) developed an estimator of the expected K-L information based on the maximized log-likelihood function. This finding makes it possible to combine ML estimation and model selection in a unified optimization framework. The criterion has a large sample formulation as:

$$\mathrm{AIC} = -2\log(L((\hat{\theta})|data)) + 2K \qquad (2a)$$

where $\log(L((\hat{\theta}|data))$ is the value of the maximized log-likelihood over the unknown parameter vector θ, given the data and the model. K is the number of estimable parameters in the model. In the special case of least-squares estimation with normally distributed errors, apart from an additive constant, AIC can be expressed as:

$$\mathrm{AIC} = n\log(\hat{\sigma}^2) + 2K \qquad (2b)$$

with

$$\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n}$$

where $\hat{\varepsilon}_i$ are the estimated residuals from the fitted model and n is the sample size. In this case the number of estimable parameters is equal to the number of model parameters plus 1 for $\hat{\sigma}^2$. In case of unequal variance at different time points, model estimates can be obtained by weighted least squares and the formula can be extended using the weighted residuals in place of $\hat{\varepsilon}_i$. Note that the additive constant depends on the number of samples, n, but not on the number of parameters, K, thereby allowing comparisons among models only for a given data set.

Assuming a set of M candidate models, AIC is computed for each model and the one with minimal AIC is selected. The selected model is also the one with minimal K-L distance from the true (unknown) data generating mechanism f. When K is large relative to n ($n/K \lesssim 40$), the small sample formulation (Hurvich and Tsai, 1989; Sugiura, 1978) should be used as:

$$\mathrm{AIC_c} = -2\log(L(\hat{\theta}|\mathrm{data})) + 2\mathrm{K} + \frac{2\mathrm{K}(\mathrm{K}+1)}{\mathrm{n\text{-}K\text{-}1}} \qquad (2c)$$

Model likelihood and Akaike weights

Define now:

$$\Delta_i = \mathrm{AIC}_i - \mathrm{minAIC} \qquad (3)$$

where minAIC is the minimum value of AIC for the model set. The likelihood of each model g_i conditional to the data and to the model set is (Akaike, 1983):

$$L(g_i|data) = \exp(-\Delta_i/2) \qquad (4)$$

Therefore, one can compare models within a set by looking at their likelihood ratios. Note that an alternative interpretation of AIC is of a maximized-likelihood

estimate where the right-hand term of Eq. 2a is a bias-correction factor due to the fact that both $\hat{\theta}$ and $L(\hat{\theta})$ are calculated from the same data set.

Since all $L(g_i|data)$ are conditional on the set of M models, it is appropriate to introduce the normalized coefficients:

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{r=1}^{M} \exp(-\Delta_r/2)} \quad (5)$$

The coefficients w_i, or "Akaike weights" (Burnham and Anderson, 1998), can therefore be interpreted as "weight of evidence" in favor of model i within the set (Akaike, 1978*a*, 1979, 1980, 1981; Buckland et al., 1997).

Akaike weights, model uncertainty, and model generalization

The Akaike weights possess an interesting interpretation in the model uncertainty framework. Suppose that each of the M models is a discrete point in a model space M. Let now the Akaike weights be ranked such that $w_1 > w_2 > \ldots > w_M$. Then the subset of space M consisting of the $T \leq M$ models such that

$$\sum_{i=1}^{T} w_i < \tau/100$$

represents the τ% confidence set conditional to the data and M (Burnham and Anderson, 1998).

This concept may be useful in the context of model selection and generalization (Breiman, 1996). In nuclear medicine, generalization refers to the common practice of selecting a model for a particular tracer on an initial data set (which we can refer to as the calibration set) and then applying this model for all subsequent studies.

The problem is that the calibration set may differ in terms of sampling, noise conditions, and parameter space from subsequent sets. For example, scan duration or time sampling may vary among studies due to technical or clinical constraints. Noise levels may also vary because of the amount of injected radioactivity, scanner type, as well as the size of the ROI. In addition, the calibration group may not be representative of a pathological population, nor indeed of the normal population (e.g., because of gender or age differences).

Therefore, unless the τ% confidence set (say, 90% confidence set) contains one model only, it may be appropriate not to discard the other models because these could provide better performance on slightly different sampling spaces.

Model averaging using Akaike weights

The concept developed in the Introduction considered a small model set where each model is a useful approximation of the data-generating process. By useful it is meant that from each model one can obtain estimations of θ_i sufficiently close to the true value of the parameter of interest, θ. The relative "closeness to the truth" of each model can be quantified in the K-L metric with AIC.

Therefore, instead of obtaining estimates of θ from the best model in the set, one can use the Akaike weights to calculate:

$$\hat{\theta} = \sum_{i=1}^{T} w_i \hat{\theta}_i \quad (6)$$

where $\hat{\theta}_i$ is the ML estimate of θ in model i (Buckland et al., 1997). Obviously, θ is assumed to be common to all models.

Note that T in Eq. 6 can be determined in two very different ways. First, one can set $T = M$ and therefore use all the models in the set. This is appropriate if M is small and the set has been defined on previous experiments. If the set is too big, however, or if some models give little or no contribution to the final estimate (i.e., their Akaike weights are generally small), then it may be appropriate to restrict the model set by setting a threshold τ as explained previously. In this case an additive uncertainty will be added owing to the model class selection procedure, but we may expect that this contribution to the total variance will be smaller than using models that have little to do with the data.

Estimates for $\mathrm{var}(\hat{\theta})$ can be obtained from linear combinations of $\mathrm{var}(\hat{\theta}_i)$ plus a component representing model misspecification bias.

Following Buckland's formalization (Buckland et al., 1997), let $\beta_i = \theta - \theta_i$ be the bias arising in estimating θ under model i. Suppose further that $E(\beta_i) = 0$ where expectation is taken over all possible models. Define:

$$E(\hat{\theta}_i|\beta_i) = \theta + \beta_i = \theta_I$$

Taking expectation over all possible models of this expression gives $E(\hat{\theta}_i) = \theta$.

Denote now: $\mathrm{var}(\hat{\theta}_i|\beta_i) = E[(\hat{\theta}_i - \theta_i)^2]$ and $\mathrm{var}(\hat{\theta}_i) = E[(\hat{\theta}_i - \theta)^2]$, then $\mathrm{var}(\hat{\theta}_i) = \mathrm{var}(\hat{\theta}_i|\beta_i) + \hat{\beta}_i^2$. This yields:

$$\mathrm{var}(\hat{\theta}) = \sum_{i=1}^{T} w_i^2 \,\mathrm{var}(\hat{\theta}_i) + \sum_{i} \sum_{l \neq i} w_i \, w_l \mathrm{cov}(\hat{\theta}_i, \hat{\theta}_l) \quad (7)$$

where the covariance term accounts for all models being estimated on the same data set. The covariance term will be in general unknown; however one can take, as an estimate, its upper bound that is equal to the geometric mean of $\mathrm{var}(\hat{\theta}_i)\mathrm{var}(\hat{\theta}_l)$ (Buckland et al., 1997). This gives:

$$\mathrm{var}(\hat{\theta}) = \{\sum_{i=1}^{T} w_i \sqrt{\mathrm{var}(\hat{\theta}|\hat{\beta}_i) + \hat{\beta}_i^2}\}^2 \quad (8)$$

Variance in Eq. 8 can be estimated by substituting $\hat{\beta} = \hat{\theta}_i - \hat{\theta}$ and $\hat{\text{var}}(\hat{\theta}_i|\beta_i) = \text{var}(\hat{\theta}_i|\beta_i)$. Alternatively, $\text{var}(\hat{\theta})$ can be obtained from Eq. 7 using bootstrap resampling (Buckland et al., 1997). See Efron and Tibshirani (1993) for an introduction to the bootstrap and Turkheimer et al. (1998) for an application of bootstrap to PET data.

Relation to other model-selection criteria

The framework developed so far does not fit the AIC criterion uniquely. Consider for example the BIC criterion (Schwarz, 1978). By defining $\Delta_i = \text{BIC}_i - \text{minBIC}$ analogously to Eq. 3, then $\exp(-\Delta_i/2)$ is equal to the posterior probability of model i conditional on model space M and the data, assuming a uniform prior on M. The crucial difference here is that BIC assumes the right model to be in M (Akaike, 1973; Buckland et al., 1997), an assumption that is believed untenable in this context. The same limitations apply to the general Bayesian model averaging approach. See Hoeting et al. (1999) for a tutorial on the subject.

Applications to compartmental modeling

Although the theory developed so far is of general use, this article focuses on the class of compartmental models used in PET for the kinetic modeling of TACs (Gunn et al., 2001; Mazoyer et al., 1986). In this context, AIC has been extensively used for model selection, and Hawkins et al. (1986) is usually the original reference for its application to PET.

The general setup for the model selection problem in PET may assume that a set M is defined consisting of a small number of physiologically meaningful compartmental models. No recipe for the definition of M is provided here. Indeed, this is not a statistical problem; model construction is a fundamental activity that is based on the existing knowledge about the system of interest (Chatfield, 1995).

In general terms, it is expected that M will contain a small number of models with a certain number of compartments. Models may not be necessarily nested, and some of models' parameters may be fixed. M may contain simultaneously models calculated over different input functions (e.g., sampled from blood, a reference region, or any other type of tomographic measurement).

Note that this framework does not apply to graphical methods like the Patlak (Gjedde, 1981; Patlak and Blasberg, 1985) or the Logan plot (Logan et al., 1990) because they are applied to reduced sets of data, the estimations are independent of the number of compartments, and, in the case of the Logan plot, they do not provide estimates in the ML sense.

Therefore, the ensuing section will evaluate the usefulness of Akaike weights in the model selection problem over such model class.

EXPERIMENTAL SECTION

The problem of model selection is common to all tracers used in PET. For the sake of clarity, this section focuses on the kinetic analysis of those radioligand studies that enable the quantitative description of neuroreceptor binding sites in the brain (Cunningham and Lammerstma, 1994; Mintun et al., 1984). A general analysis of compartmental models in PET is presented by Gunn et al. (2001), where a distinction is made between "macro" parameters (e.g., the total volume of distribution), which have the same interpretation across a set of particular compartmental models, and "micro" parameters, the interpretation of which is model specific. In this context, we consider the general compartmental model illustrated diagrammatically in Fig. 1A. This model represents the possible exchanges of the tracer between a vascular (plasma) compartment and three tissue compartments: a "free" pool, a "non–specifically bound" pool, and a "specifically bound" pool. The corresponding rate equations for this model are as follows:

$$C_{tot}(t) = C_f(t) + C_b(t) + C_{ns}(t)$$
$$dC_f(t)/dt = K_1C_a(t) - (k_2 + k_3 + k_5)\,C_f(t) + k_4C_b(t) + k_6C_{ns}(t)$$
$$dC_b(t)/dt = k_3C_f(t) - k_4C_b(t)$$
$$dC_{ns}(t)/dt = k_5C_f(t) - k_6C_{ns}(t).$$

$C_{tot}(t)$ denotes the total concentration of radioligand in the tissue as a function of time, comprising free ($C_f(t)$), non–specifically bound ($C_{ns}(t)$), and specifically bound ($C_b(t)$). It is here assumed that measures of the concentration of label in arterial whole blood [$C_{wb}(t)$] and of the parent radioligand in

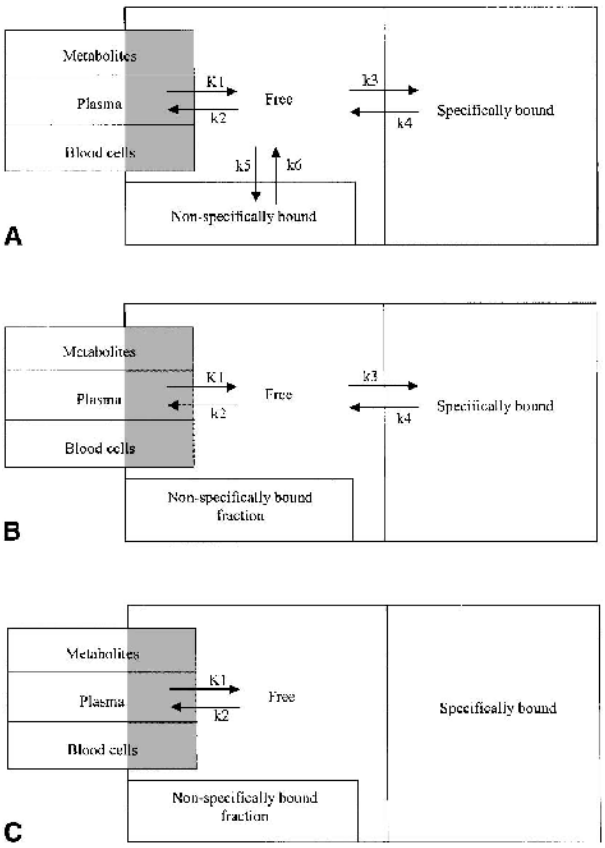

FIG. 1. (A–C) One-compartment model.

TABLE 1. *Akaike information coefficient weights for [^{11}C]flumazenil models*

	Model		
	A	C	E
Frontal cortex	0.03	0.10	0.87
Occipital cortex	0.17	0.55	0.28
Caudate	0.07	0.39	0.54
Putamen	0.00	0.01	0.99
Thalamus	0.45	0.41	0.14
Cerebellum	0.00	0.01	0.99
Pons	0.00	0.79	0.21

arterial plasma [input function $C_a(t)$] are available. It is also assumed that radiolabeled metabolites of the parent in plasma do not cross the blood–brain barrier.

K_1 and $k_2 - k_6$ are the rate constants; K_1 is the unidirectional clearance of the ligand from plasma to tissue whereas k_2 models the clearance from tissue to plasma; k_5 and k_6 model the exchanges of the free tissue compartment with the nonspecific binding space; $k_3 = K_{on} B_{avail}$, where K_{on} is the bimolecular association rate constant and B_{avail} is the concentration of available binding sites; and $k_4 = K_{off}$ is the dissociation rate constant of the ligand-binding site complex. Finally, because of the limited spatial resolution of PET, the total PET signal comprises also label in the vasculature. Denoting the vascular volume fraction as V_b yields:

$$C_{PET}(t) = (1 - V_b)C_{tot}(t) + V_b C_{wb}(t)$$

The parameter of interest in this model is usually the ratio k_3/k_4 termed "binding potential" (Mintun et al., 1984). A simplification often used is the description of the radioligand uptake in terms of volume of distribution. The total volume of distribution (VD_{tot}) is the partition ratio between the total concentration of ligand in the tissue and that in plasma that would be attained in a steady-state equilibrium between the two. The estimate of this parameter is generally more stable than the estimates of microparameters (K_1, k_2- k_6) and can be calculated as:

$$VD_{tot} = K_1/k_2(1 + k_3/k_4 + k_5/k_6) \quad (9a)$$

The three compartments model can be simplified to two compartments (Fig. 1B) by assuming a fast equilibration between the free and the nonspecific fractions in the tissue (Koeppe et al., 1991). In this case, the volume of distribution is calculated as:

$$VD_{tot} = K_1/k_2(1 + k_3/k_4) \quad (9b)$$

If tissue compartments are indistinguishable than the model reduces to one compartment only (Fig. 1C). Concurrently VD_{tot} becomes:

$$VD_{tot} = K_1/k_2 \quad (9c)$$

The compartment model above is here defined for illustrative purposes only. There is no suggestion that this model is generally "right" or that it is the most practical.

Measured data set: modeling ligand binding to the GABA receptor

For several radioligands, the difference in fit quality between models is small (Carson et al., 1998; Lammertsma et al., 1996). As an example, this section focuses on the kinetic modeling of ligands binding to the central benzodiazepine (GABA) receptor in the brain. Consider the data of Koeppe et al. (1991) on the compartmental modeling of [^{11}C]flumazenil, a high-affinity antagonist for GABA receptors.

In this study, three models were considered. The first, the two-parameter model, is shown in Fig. 1C. The second, the three-parameter model, is the two-compartment model shown in Fig. 1B where the ratio K_1/k_2 was fixed to a value estimated with an ROI placed on the pons. The third is the four-parameter model shown in Fig. 1B with all parameters to be estimated. Blood volume, V_b, was inserted in all three models as estimated from the entire brain volume. For reasons that will be clearer later, we label these models respectively as model A, model C, and model E (Table 3).

For each model, the χ^2 reported by Koeppe et al. (1991) for six ROIs averaged over six subjects were multiplied the degrees of freedom, e.g.,

$$\chi^2 = \left(\frac{\sum \varepsilon_i^2}{n - K}\right)$$

to obtain the sum of the squared residuals. The Akaike weights were then calculated as in Eq. 5 using the AICc of Eq. 2c and are shown in Table 1. Note that, since the χ^2 were averaged, these weights are not mean weights, but weights relative to the mean likelihood.

For the values of the estimated parameters the reader is referred to Table 2 where values are reported for use in a simulation study, or to the original reference. Analysis of Table 1 suggests that model E is selected with little uncertainty in half of the regions (frontal cortex, putamen, cerebellum). Selection is far less clear-cut in the other cases. In the thalamus, for example, all three models considered offer similar likelihoods.

The data of Buck et al. (1996) provide further evidence that the problem relates more generally to the system being modeled and it is not peculiar of a single tracer or data set. In this case, the GABA receptor density was measured using [^{11}C]iomazenil and a similar array of models was used. The data analysis considered the two-parameter model (model A following Buck's labeling), the four-parameter model (model E) and three subversions of it where either K_1/k_2 (model C) or k_4 (model D) or both (model B) were fixed to preestimated values. These models are summarized in Table 3. The AICc values for

TABLE 2. *Kinetic parameters for [^{11}C]flumazenil simulation*

	K_1	k_2	k_3	k_4	k_5	k_6	VD_{tot}
Frontal cortex	0.325	0.113	0.55	0.28	0.04	0.12	9.53
Occipital cortex	0.335	0.010	0.62	0.40	0.04	0.12	9.69
Caudate	0.319	0.175	0.12	0.18	0.04	0.12	3.64
Putamen	0.359	0.156	0.13	0.11	0.04	0.12	5.78
Thalamus	0.384	0.215	0.59	0.49	0.04	0.12	5.03
Cerebellum	0.314	0.186	0.31	0.20	0.04	0.12	4.87

Values of K_1 are given in mL · g^{-1} · min^{-1}, other constants are in min^{-1}; VD_{tot} is dimensionless and is calculated according to Eq. 9a.

TABLE 3. *[^{11}C]Iomazenil models*

Model	No. compartments	No. free parameters	Fixed parameters
A	2	2	—
B	3	2	K_1/k_2, k_4
C	3	3	K_1/k_2
D	3	3	k_4
E	3	4	—

the five models are reported in Table 4 and were calculated for two regions (cerebellum and frontal cortex) and averaged over five subjects. The values were obtained from Table 2 of the original reference.

Table 4 provides a picture very similar to the one provided by Table 1. In the case of the occipital cortex, the two-compartment model in its various arrangements (B, C, D, and E) provides similar likelihood. Note that although these models arise from the same compartmental structure, they provide numerical estimates of the parameters that differed markedly (Buck et al., 1996).

It is also remarkable that a region with low binding (i.e., the cerebellum) has consistently better fit with the maximum number of free parameters (model E). This observation echoes similar ones made in previous reports (Carson et al., 1998; Watabe et al., 2000) on the revealing fact that regions with little or no specific binding often require the most complex model. Most probably, the absence of a relatively large kinetic component due to specific uptake brings to the identifiability level a larger number of compartments, of comparable magnitude, that are involved in the nonspecific binding of the ligand. Alternatively, reference regions may be bigger than regions with specific binding and therefore possess a more favorable signal-to-noise ratio.

Synthetic data set: simulated [^{11}C]flumazenil study

The information acquired from the experimental data can now be used to attempt a numerical evaluation of the weighted average technique. The artificial data sets considered a "reality" similar to the one of the real data sets on [^{11}C]flumazenil. Note that by "similar" it is not meant that we attempt to duplicate reality: this is not possible and, as it will be clear from the results, matching results between simulations and reality have not been obtained. Instead a scenario is created with useful similarities to these data.

The kinetic of [^{11}C]flumazenil is simulated using the three-compartment, six-parameter model of Fig. 1A for six regions. Kinetic parameters for the regions were obtained from Koeppe et al. (1991) and are illustrated in Table 2. The constants for the nonspecific compartment (k_5, k_6) were obtained from k_3, k_4 values of the pons, where no specific binding is expected.

An artificial noiseless arterial input function resembling real ones was used to generate the TACs that consisted in 21 frames for a total of 60 minutes. Gaussian noise was then added to TACs with standard deviation proportional to the absolute activity in each frame. The noise proportionality constant varied from 5% for large regions (frontal and occipital cortex, cerebellum) to 7% or 8% for the smaller ones (caudate, putamen, thalamus) to match real conditions.

TABLE 4. *Akaike information coefficients weights for [^{11}C]iomazenil models*

	Model				
	A	B	C	D	E
Cerebellum	0.00	0.04	0.04	0.17	0.75
Occipital cortex	0.00	0.19	0.32	0.31	0.18

The simulation considered the estimation of the same three models used by Koeppe et al. (1991) and described in the previous section. Note that none of these models considers the additional third compartment simulating nonspecific binding of the tracer. This was arranged with the purpose of having the supposedly "true" model outside the model set. Indeed this situation may be quite realistic given that many tracers bind to nonspecific sites and PET data do not usually allow the estimation of models with more than three compartments.

Model parameters were estimated using a nonlinear least-square procedure performed using the Nelder-Mead simplex minimization routine (Nelder and Mead, 1965). Software was written in Matlab (The Mathworks Inc., MA, U.S.A.) and run on an Ultra-Sparc 10 Sun Workstation (SunSystems, Mountain View, CA, U.S.A.). Volumes of distribution were calculated according to Eq. 9c (model A) and Eq. 9b (models C and E).

The first set of results contemplates an overall numerical comparison between the model averaging and model selection methods. Two hundred artificial data sets, each with six regional TACs, were generated and the three models estimated. For each TAC, AICc coefficients for the three models were calculated. The parameter of interest, in this case the total volume of distribution VD_{tot}, was calculated either from the model with lowest AICc (model selection) or as a weighted average from the three models (as in Eq. 6) using the Akaike weights.

RESULTS

The results of the simulation are given in Tables 5, 6 and 7 in terms of percentage bias, standard deviation, and MSE over the estimated VD_{tot}. Table 8 contains the frequencies of selection for each model by the minimum AICc procedure for the six regions. Table 9 shows the average weights for the AICc-weighted procedure.

Tables 5 and 6 and particularly the MSE values contained in Table 7 suggest that model E, the most complex model, performs better in the regions with lower specific uptake (i.e., the caudate and putamen), where the relative effect of the nonspecific component is greater. Instead, model C has slightly better results on the remaining regions. MSE values for model A are the highest for the three models in all regions and are reasonable only on the regions with high specific uptake (i.e., the cortex).

The model averaging method shows a general improvement of the MSE compared with that of the model selection that ranged from 21% to 10% in the low-uptake regions (caudate and putamen) and 8% to 4% in the remaining regions. Note in Table 9 how the AICc-weighted procedure balances all three models' contributions to obtain VD_{tot} estimates.

Synthetic data set: simulated [^{11}C]flumazenil study revisited

The simulation results obtained so far can be used to deepen our analysis of the modeling process. Now con-

TABLE 5. *Volume of distribution: bias for [^{11}C]flumazenil simulation*

	Model			Model averaging	Model selection
	A	C	E		
Frontal cortex	−1.59	−0.48	−0.56	−0.93	−1.01
Occipital cortex	−1.44	−0.73	−0.88	−1.14	−1.25
Caudate	−5.97	0.84	−0.07	0.01	0.22
Putamen	−7.26	1.67	0.67	1.15	1.33
Thalamus	−2.32	−0.82	−0.94	−1.39	−1.45
Cerebellum	−3.60	0.01	−0.49	−0.45	−0.26

Values are given in percentage.

TABLE 6. *Volume of distribution: standard deviation for [^{11}C]flumazenil simulation*

	Model			Model averaging	Model selection
	A	C	E		
Frontal cortex	2.64	2.50	2.47	2.49	2.57
Occipital cortex	2.83	2.68	2.77	2.76	2.77
Caudate	2.84	3.01	2.52	2.67	2.93
Putamen	2.26	2.66	2.26	2.40	2.57
Thalamus	3.07	2.84	2.82	2.81	2.97
Cerebellum	2.42	2.19	2.30	2.12	2.24

Values are given in percentage over the mean value.

TABLE 7. *Volume of distribution: mean squared error for [^{11}C]flumazenil simulation*

	Model			Model averaging	Model selection
	A	C	E		
Frontal cortex	9.44	6.43	6.35	7.02	7.55
Occipital cortex	9.99	7.62	8.39	8.82	9.15
Caudate	43.65	9.66	6.30	7.05	8.53
Putamen	57.79	9.82	5.51	7.03	8.31
Thalamus	14.73	8.67	8.75	9.77	10.83
Cerebellum	18.78	4.77	5.46	4.63	5.04

Bias squared plus variance.

sider that the previous comparison does not mimic exactly the way model selection is carried out in practice. Model selection is usually not carried out on all data sets but on a subset (e.g., 10 to 15 subjects), and the model selected in this subset is then used on the following data. Besides, one model and one model only is usually selected for use on all the ROIs (see Koeppe et al., 1991).

Therefore, to better simulate reality, the results of the previous simulation should be used in a different way. The model selection procedure shall now consist in using the results of the first 12 data sets (this is a common sample size in PET) to choose the model, and use this model on all ROIs for the remaining data sets.

In this case, the results of the first 12 simulations matched the ones of Table 8. Model A (two-compartment, two-parameter model) was the one more consistently selected for the high-uptake regions. Therefore, following real practice (Koeppe et al., 1991), model A should also be selected for all the other ROIs. This implies that the results of the "real practice" model selection procedure are the ones in the first column, those corresponding to model A. In this case, the improvement in MSE of the model averaging procedure toward a "realistic" model selection approach ranges from 10% to 30% in the cortical regions to 400% to 800% in the smallest ones (caudate and putamen).

Synthetic data set: simulated [^{11}C]flumazenil and generalization error

As explained previously, the generalization error is that error incurred by an estimation procedure when it is used on a sample space that is different from the sample space where it has been optimized. In our simulation, a traditional model selection procedure chose model A as optimal over the data set derived from the "normal" set of kinetic constants shown in Table 4. Now we construct a reasonable "pathologic" set by simulating a reduction in specific binding, manufactured with a 20% reduction of k_3 in all ROIs. This new set of constants is used to generate 200 artificial data sets in the same fashion as before.

The results of this simulation are shown in the usual terms of bias, standard deviation, and MSE in Tables 10, 11, and 12. Table 13 shows frequencies of models selected by the minimum Akaike procedure. Table 14 shows the average weights for the AICc-weighted procedure.

TABLE 8. *Frequencies of model selected by minimum Akaike information coefficient criterion for [^{11}C]flumazenil simulation*

	Model		
	A	C	E
Frontal cortex	0.54	0.43	0.03
Occipital cortex	0.79	0.18	0.03
Caudate	0.12	0.74	0.14
Putamen	0.01	0.72	0.27
Thalamus	0.46	0.48	0.06
Cerebellum	0.12	0.86	0.02

TABLE 9. *Mean values for Akaike weights for [^{11}C]flumazenil simulation*

	Model		
	A	C	E
Frontal cortex	0.49	0.41	0.10
Occipital cortex	0.60	0.32	0.08
Caudate	0.12	0.61	0.27
Putamen	0.01	0.62	0.37
Thalamus	0.42	0.44	0.14
Cerebellum	0.13	0.70	0.17

TABLE 10. *Volume of distribution: bias for [^{11}C]flumazenil simulation, pathologic set*

	Model			Model averaging	Model selection
	A	C	E		
Frontal cortex	−1.95	−0.41	−0.64	−0.94	−0.93
Occipital cortex	−1.81	−0.87	−0.89	−1.15	−1.21
Caudate	−6.92	0.78	−0.33	0.02	0.05
Putamen	−8.21	1.67	0.48	0.97	0.99
Thalamus	−3.22	−1.50	−1.47	−1.84	−1.79
Cerebellum	−4.53	0.03	−0.68	−0.45	−0.28

Values are given in percentage.

TABLE 11. *Volume of distribution: standard deviation for [^{11}C]flumazenil simulation, pathologic set*

	Model			Model averaging	Model selection
	A	C	E		
Frontal cortex	2.47	2.41	2.38	2.43	2.56
Occipital cortex	2.61	2.50	2.61	2.55	2.63
Caudate	3.09	2.85	2.42	2.68	3.02
Putamen	2.22	2.51	2.13	2.20	2.35
Thalamus	2.98	2.93	3.02	2.81	2.83
Cerebellum	2.54	2.40	2.40	2.28	2.50

Values are given in percentage over the mean value.

TABLE 12. *Volume of distribution: mean squared error for [^{11}C]flumazenil simulation, pathologic set*

	Model			Model averaging	Model selection
	A	C	E		
Frontal cortex	19.76	11.86	12.06	13.48	14.80
Occipital cortex	20.07	13.96	15.17	15.54	16.71
Caudate	114.84	17.36	11.86	14.44	18.15
Putamen	144.56	21.58	9.50	11.51	12.99
Thalamus	38.43	21.58	22.45	22.42	22.52
Cerebellum	53.88	11.51	12.43	10.79	12.61

Bias squared plus variance.

Table 12 shows that the reduction in specific uptake induced greater MSE owing to the relative increase of the nonspecific compartment contribution that is not considered by any of the three models used. As before, model averaging performed better than model selection in MSE terms over all regions. Table 14 illustrates how weights in the more complex models, C and E, have gradually increased to compensate for the change in kinetic properties of the system. Remember, however, that the model selected in the "normal" set was model A. Model A performance on the "pathologic set" is obviously degraded with an almost doubled increase in MSE compared with the model averaging method. Note that on this set, the model selected by the minimum AIC procedure was model C also for the frontal cortex (see Table 13).

DISCUSSION

The results contained in the experimental section, for real or simulated data, allow the following inferences. First, reality is complex and although no optimal model exists, better ones do. The distance of a model from the true data-generating mechanism can be calculated from its mean likelihood, and AIC coefficients are good estimates of this distance. Second, there may be instances where different models in a set may have comparable likelihood over the same data set. In this case, rejection of a model may cause more harm than good, and better estimates can be obtained by averaging the parameter of interest over the entire model set. The limitation is that all models must allow the estimation of the parameter of interest. Third, rejection of models can be even more harmful when one considers the generalization error, the error incurred when the model selected is applied on different data sets. Obviously, there will also be instances where model uncertainty will be low, and in this case model averaging will be reduced to the use of a single robust model. Finally, a method that allows the use of an array of models is a strong conceptual tool. For example, when physiologically appropriate, one could conceive the simultaneous use of blood and reference inputs to improve the statistical properties of the estimates. Furthermore, one could adapt more efficiently model complexity to varying spatial scales. This concept applies to the varying size of ROIs, as illustrated in the experimental section, but it could be also extended to

TABLE 13. *Frequencies of models selected by minimum Akaike information coefficient criterion for [^{11}C]flumazenil simulation, pathologic set*

	Model		
	A	C	E
Frontal cortex	0.45	0.53	0.02
Occipital cortex	0.71	0.23	0.06
Caudate	0.12	0.75	0.13
Putamen	0	0.68	0.32
Thalamus	0.37	0.54	0.09
Cerebellum	0.09	0.85	0.06

TABLE 14. *Mean values for Akaike weights for [^{11}C]flumazenil simulation, pathologic set*

	Model		
	A	C	E
Frontal cortex	0.42	0.48	0.10
Occipital cortex	0.56	0.33	0.11
Caudate	0.12	0.60	0.28
Putamen	0	0.62	0.38
Thalamus	0.36	0.46	0.18
Cerebellum	0.10	0.70	0.20

multiscale approaches for the production of parametric maps (Turkheimer et al., 2000). These are among the possibilities that will be explored in future work.

Acknowledgments: The authors would like to thank Kathy Schmidt (NIMH, Bethesda, MD, U.S.A.) for reading the manuscript and providing useful comments.

REFERENCES

Akaike H (1973) Information theory and extension of the maximum likelihood principle. In: *2nd International Symposium in Information Theory* (Petrov BN, Csaki F, eds), Budapest, Hungary: Akademiai Kiado, pp 267–281

Akaike H (1978*a*) A Bayesian analysis of the minimum AIC procedure. *Ann Inst Stat Math* 30:9–14

Akaike H (1978*b*) On newer statistical approaches to parameter estimation and structure determination. *Int Fed Automatic Control* 3:1877–1884

Akaike H (1979) A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika* 66:237–242

Akaike H (1980) Likelihood and the Bayes procedure (with discussion). In: *Bayesian statistics* (Bernardo JM, De Groot MH, Lindley DV, Smith AFM, eds), Valencia, Spain: University Press, pp 143–203

Akaike H (1981) Modern development of statistical methods. In: *Trends and progress in system identification* (Eykhoff P, ed), Paris, France: Pergamon Press, pp 169–184

Akaike H (1983) Information measures and model selection. *Int Stat Inst* 44:277–291

Allen DM (1974) The relationship between variable selection and data augmentation and a method of prediction. *Technometrics* 22 :125–127

Boltzmann L (1877) Ueber die Beziehung zwischen dem Haupsatze der mechanischen Waermetheorie und der Wahrscheinlichkeitsrechnung respective den Saetzen ueber das Waermegleichgewicht. *Wiener Berichte* 76:373–345

Breiman L (1996) Heuristics of instability and stabilization in model selection. *Ann Stat* 24:2350–2383

Buck A, Westera G, vonSchulthess GK, Burger C (1996) Modeling alternatives for cerebral carbon-11-iomazenil kinetics. *J Nucl Med* 37:699–705

Buckland ST, Burnham KP, Augustin NH (1997) Model selection: an integral part of inference. *Biometrics* 53:603–618

Burnham KP, Anderson DR (1998) *Model selection and inference: a practical information-theoretic approach* New York, NY: Springer-Verlag

Carson RE, Kiesewetter DO, Jagoda E, Der MG, Herscovitch P, Eckelman WC (1998) Muscarinic cholinergic receptor measurement with [^{18}F]FP-TZTP: control and competition studies. *J Cereb Blood Flow Metab* 18:1130–1142

Chatfield C (1995) Model uncertainty, data mining and statistical inference. *J R Stat Soc Series A* 158:419–466

Cunningham VJ, Lammerstma AA (1994) Radioligand studies in brain: kinetic analysis of PET data. *Med Chem Res* 5:79–96

Efron B, Tibshirani RJ (1993) *An introduction to the bootstrap* New York, NY: Chapman and Hall

Gjedde A (1981) High- and low- affinity transport of D-glucose from blood to brain. *J Neurochem* 36:1436–1471

Gunn RN, Gunn SR, Cunningham VJ (2001) PET compartmental models. *J Cereb Blood Flow Metab* 21:279–287

Hawkins RA, Phelps ME, Huang S-C (1986) Effects of temporal sampling, glucose metabolic rates, and disruption of the blood-brain barrier on the FDG model with and without a vascular compartment: studies in human brain tumors with PET. *J Cereb Blood Flow Metab* 6:170–183

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. *Stat Sci* 14:382–417

Hurvich CM, Tsai C-L (1989) Regression and time series model selection in small samples. *Biometrika* 76:297–307

Koeppe RA, Holthoff VA, Frey KA, Kilbourn MR, Kuhl DE (1991) Compartmental analysis of [^{11}]flumazenil kinetics for the estimation of ligand transport rate and receptor distribution using positron emission tomography. *J Cereb Blood Flow Metab* 11:735–744

Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* 22:79–86

Lammertsma AA, Bench CJ, Hume SP, Osman S, Gunn RN, Brooks DJ, Frackowiack RSJ (1996) Comparison of methods for analysis of [C-11]raclopride studies. *J Cereb Blood Flow Metab* 16:42–52

Logan J, Fowler JS, Volkow ND, Wolf AP, Dewey SL, Schlyer DJ, Macgregor RR, Hitzmann R, Bendriem B, Gatley SJ, Christman DR (1990) Graphical analysis of reversible radioligand binding from time-activity measurements applied to [*N*-^{11}C-methyl]-(-)-cocaine PET studies in human subjects. *J Cereb Blood Flow Metab* 10:740–747

Mallows C (1973) Some comments on Cp. *Technometrics* 15:661–675

Mazoyer BM, Huesman RH, Budinger TF (1986) Dynamic PET data analysis. *J Comput Assist Tomogr* 10:645–653

Mintun MA, Raichle ME, Kilbourn MR, Wooten GF, Welch MJ (1984) A quantitative model for the *in vivo* assessment of drug binding sites with positron emission tomography. *Ann Neurol* 15:217–227

Nelder JA, Mead R (1965) A simplex method for function minimization. *Comput J* 7:308–313

Patlak CS, Blasberg RG (1985) Graphical evaluation of blood-to-brain transfer constants from multiple-time uptake data: generalizations. *J Cereb Blood Flow Metab* 5:584–590

Rissanen J (1978) Modeling by shortest data description. *Automatica* 14:465–471

Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464

Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423, 623–656

Soofi ES (1994) Capturing the intangible concept of information. *J Am Stat Assoc* 89:1243–1254

Stone M (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J R Stat Soc Series B* 36:111–147

Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun Stat Theory Meth* A7:13–26

Turkheimer FE, Banati RB, Visvikis D, Aston JAD, Gunn RN, Cunningham VJ (2000) Modeling dynamic PET-SPECT studies in the wavelet domain. *J Cereb Blood Flow Metab* 20:879–893

Turkheimer FE, Sokoloff L, Bertoldo A, Lucignani G, Reivich M, Jaggi JL, Schmidt K (1998) Estimation of component and parameter distributions in spectral analysis. *J Cereb Blood Flow Metab* 18:1211–1222

Watabe H, Channing MA, Der MG, Adams R, Jagoda E, Herscovitch P, Eckelman WC, Carson RE (2000) Kinetic analysis of the 5-HT$_{2A}$ ligand [^{11}C]MDL100,907. *J Cereb Blood Flow Metab* 20:899–909