Variability and Uncertainty of Biokinetic Model Parameters: The Discrete Empirical Bayes Approximation

Guthrie Miller*
Los Alamos National Laboratory, Los Alamos, NM, USA

Received March 3 2008, revised May 21 2008, accepted June 10 2008

In the Bayesian approach to internal dosimetry, uncertainty and variability of biokinetic model parameters need to be taken into account. The discrete empirical Bayes approximation replaces integration over biokinetic model parameters by discrete summation in the evaluation of Bayesian posterior averages using Bayes theorem. The discrete choices of parameters are taken as best-fit point determinations of model parameters for a study subpopulation with extensive data. A simple heuristic model is constructed to numerically and theoretically study this approximation. The heuristic example is the measurement of heights of a group of people, say from a photograph where measurement uncertainty is significant. A comparison is made of posterior mean and standard deviation of height after a measurement, (i) using the exact prior describing the distribution of true height in the population and (ii) using the approximate discrete empirical Bayes prior obtained from measurements of some study subpopulation.

INTRODUCTION

Internal dosimetry relies on biokinetic models to relate the measured bioassay quantities, for example, urinary excretion, to the imparted internal dose. It is well recognised that biokinetic model parameters are uncertain and variable in any population of interest. In the Bayesian approach to internal dosimetry\(^1\), biokinetic model parameter variability and uncertainty can be taken into account by averaging over a discrete set of biokinetic models (different choices of model parameters). This assertion follows from the purely mathematical fact that the integration over biokinetic model parameters that appears in Bayes theorem can be approximated by a discrete summation. For this approach to be useful in practice, one must decide how to choose the finite discrete set of biokinetic models and have some idea of the errors introduced relative to exact evaluation of the Bayesian integrals.

One always seeks to determine the biokinetic prior probability distribution as much as possible from measurement data. Such data usually take the form of a representative study subpopulation with extensive high quality measurements. How then is one to use such a study data set to determine a biokinetic prior? A number of approaches are currently being studied\(^2,3\). The approach described here is quite simple and intuitive. Point determinations of biokinetic parameters are made for each case in the study subpopulation using minimum-\(\chi^2\) data fitting\(^4\), where certain key biokinetic parameters are chosen to be variable. These point determinations are the discrete models that are to be averaged over in the evaluation of posterior probabilities using Bayes theorem. A generalisation of this approach\(^4\) not discussed further here would be to generate some number (\(\gg 1\), perhaps on the order of 10–100) alternate realisations of the biokinetic parameters for each study case to represent uncertainty.

As an example, for plutonium using standard International Commission on Radiation Protection (ICRP) models, the number of biokinetic parameters is on the order of 50. Even a study subpopulation with extensive good data is usually not sufficient to determine all these parameters, so assumptions have to be made about which parameters are of most importance. The problem is simplified by allowing variations only of these key parameters. The ultimate Bayesian method would need to allow all parameters to vary and, using the data from the study subpopulation, would determine their joint posterior distribution, which is to be used as a prior for other cases. This ultimate statistical approach is many years if not decades off in the future for cases with many parameters, like plutonium using standard ICRP models.

This paper considers a simple heuristic example, the measurement of heights of persons in a particular population. This example is sufficiently simple so...
that calculations are easily carried out, yet it is still instructive.

THE DISCRETE EMPIRICAL BAYES METHOD

The discrete empirical Bayes method uses point determinations of biokinetic parameters from a representative set of study subpopulations with good data from a population to construct a prior probability distribution of biokinetic parameters for this population. There are two sources of uncertainty: (1) inter-individual variability in the population and (ii) measurement uncertainty. The simple empirical Bayes method assumes that variability dominates measurement uncertainty (for the measurements used to determine the prior). Because the sample of cases is considered to be representative of the population being studied as a whole, the empirical-Bayes prior would be an inter-individual variability mixture of distributions, each component of the mixture corresponding to one case, for a large sample of representative cases. There are imagined to be two types of measurements: (i) prior-determination measurements of high quality (quality perhaps representing quantity of data for a real-life study subpopulation) and (ii) the normal measurements that are to be used with the empirical Bayes prior to determine the quantities of interest. A particular study case determines a single prior mixture component, which is representative set of study subpopulations with good data from a population to construct a prior probability distribution of biokinetic parameters for this population.

the minimum-deviation for each study case with a delta-function Bayes method merely replaces the posterior distribution resulting from the prior-determination measurements for that case. This discrete empirical Bayes method merely replaces the posterior distribution for each study case with a delta-function distribution corresponding to the minimum-\( \chi^2 \) determination of biokinetics parameters for that case.

As a simple heuristic example of the discrete empirical Bayes method, it is imagined that it is desired to determine the heights of a group of persons in a situation where measurement uncertainty is significant, say when heights are determined from a photograph. In the case of heights of persons, there are readily available data\(^5\) that could be used to construct a prior probability distribution of heights based on knowledge of the population (e.g. age, sex and ethnicity). Thus, it is imagined that there is a known, exactly correct prior. This prior distribution of heights in the population is assumed to be a Gaussian distribution with mean value \( h_0 \) and standard deviation \( \sigma_0 \). The discrete empirical prior will be compared with this exact prior.

For the determination of the discrete empirical Bayes prior, it is assumed that the measurement technique is known to have a Gaussian likelihood function with standard deviation \( \sigma \). To test the discrete empirical Bayes method, a single measurement with result \( M \) and Gaussian-likelihood standard deviation \( \sigma_m \) is interpreted using (i) the exact formula from Bayes theorem, and (ii) the discrete empirical Bayes approximation based on \( N \) cases used to determine the prior, and the two interpretations compared.

When the distribution of true heights in the population \( P(h) \) is known exactly and this is used as the prior, Bayes theorem\(^6\) gives the posterior probability distribution of true height after the measurement as

\[
P(h|M) = \frac{P(M|h)P(h)}{\int P(M|h)P(h)dh},
\]

where \( P(M|h) \) is the probability of obtaining measurement result \( M \) when the true height is \( h \), assumed to be given by

\[
P(M|h) = \frac{1}{\sqrt{2\pi \sigma_m^2}} \exp\left(-\frac{1}{2} \left( \frac{M - h}{\sigma_m} \right)^2 \right).
\]

In Bayesian inference, the measurement result \( M \) is known exactly and this is used as the prior, Bayes theorem, equation (1).

Because both the prior and the likelihood function in equation (1) are Gaussian, the posterior given by equation (1) is also Gaussian with mean value (the exact posterior mean height)

\[
h^{(\text{exact})} = \frac{h_0/\sigma_0^2 + M/\sigma_m^2}{1/\sigma_0^2 + 1/\sigma_m^2},
\]

and posterior standard deviation

\[
\sigma_h^{(\text{exact})} = \sqrt{\frac{1}{1/\sigma_0^2 + 1/\sigma_m^2}}.
\]

Figure 1 illustrates the situation. The posterior mode (maximum probability point) is pulled away from the likelihood function mode in the direction of the prior mode.

The corresponding plot in terms of cumulative probability is shown in Figure 2.

The discrete empirical Bayes method assumes for the prior a mixture of delta-function distributions at the measured heights \( h_i \) for \( i = 1 \ldots N \), where \( N \) is the number of cases in the study subpopulation used. This is illustrated in Figure 3 for \( N = 30 \).

In the discrete empirical Bayes approximation, the posterior average of a general function \( f \) of true height \( h \) is given by

\[
< f > = \int f(h)P(h|M)dh = \frac{\sum_{i=1}^{N} f(h_i)P(M|h_i)}{\sum_{i=1}^{N} P(M|h_i)},
\]

where \( f(h_i) \) is the function of the measured heights and \( P(M|h_i) \) is the likelihood of the measured data from each case.

\( \Delta V \) and \( \Delta \sigma \) for the total uncertainty are calculated as

\[
\Delta V = \frac{\sum_{i=1}^{N} \left( h_i - h^{(\text{exact})} \right)^2 \sigma_i^2}{\sum_{i=1}^{N} \sigma_i^2},
\]

and

\[
\Delta \sigma = \sqrt{\frac{\sum_{i=1}^{N} \left( h_i - h^{(\text{exact})} \right)^2 \sigma_i^2}{\sum_{i=1}^{N} \sigma_i^2} - \left( \frac{\sum_{i=1}^{N} \left( h_i - h^{(\text{exact})} \right)^2 \sigma_i^2}{\sum_{i=1}^{N} \sigma_i^2} \right)^2},
\]

where \( \sigma_i \) is the standard deviation of the measurement uncertainty at the height \( h_i \).

If \( \sigma_i \) is constant across all cases, \( \Delta V \) and \( \Delta \sigma \) are proportional to the number of cases.

\( \sigma_0 \) and \( \sigma_1 \) are the standard deviations of prior and likelihood, respectively.
where $h_i$ is the $i$th measurement result of the $N$ measurements used to determine the prior. The quantity $h_i$ corresponds in this simple example to the minimum-$\chi^2$ point determination spoken of above. It is a single best value rather than a distribution. The measured heights in equation (5) are those that would occur when one randomly selects $N$ cases from the population and measures their heights. For the numerical study, these height measurement results are generated from a Gaussian distribution with mean $h_0$ and standard deviation $\sqrt{\sigma_0^2 + \sigma^2}$, where the measurements used to determine the prior are assumed to have standard deviation $\sigma$; that is, the distribution of height measurements is broader than the distribution of true values of height, being affected by measurement uncertainty as well as being a measure of variability in the population. In equation (5), $P(M|h_i)$ is the likelihood function corresponding to the $i$th study case, given by equation (2). Equation (5) can be used to obtain the posterior mean and standard deviation in the discrete empirical Bayes approximation by letting the function $f(h) = h$ and $f(h) = h^2$.

A standard diagnostic is $\chi^2$, given in this case by

$$\chi^2 = \left( \frac{M - <h>}{\sigma_m} \right)^2.$$  \hspace{1cm} (6)

Values of $\chi^2 >> 1$ indicate that more mixture components are needed in the prior, e.g. when the measurement has small uncertainty, and the prior is very ‘grainy’. This is illustrated in Figure 4.

Figure 5 shows the difference of the posterior means calculated using the exact prior and the discrete empirical Bayes prior, normalised to the exact posterior standard deviation as a function of the number $N$ of measurements determining the discrete prior (the measurement value $M$ is assumed to be $h_0 - 1/2 \sigma_0$).

For this numerical example, measurement uncertainty is assumed to be larger than inter-individual variability in the population. From Figure 5, the difference in the posterior means calculated in the two ways is small compared with the posterior standard deviation. This shows that the discrete empirical Bayes approximation is adequate. Also, one sees that measurement uncertainty for the measurements used to determine the empirical Bayes prior is unimportant in this situation.

Figure 6 shows the ratios of the posterior standard deviations calculated in the two ways as a function
of the number \( N \) of mixture components of the discrete prior.

One can show that the ratio of standard deviations in Figure 6 tends to

\[
\sqrt{\frac{1/\sigma_0^2 + 1/\sigma_m^2}{1/(\sigma_0^2 + \sigma^2) + 1/\sigma_m^2}} = 1.22. \tag{7}
\]

Thus, the discrete empirical Bayes posterior standard deviation is somewhat larger than it should be. This error goes away if the measurements used to determine the prior have small uncertainty compared with variability in the population (\( \sigma \) much less than \( \sigma_0 \)) or if measurement uncertainty is much smaller than population variability (\( \sigma_m \) is much smaller than \( \sigma_0 \)).

Figure 7 shows \( \chi^2 \) as a function of the number \( N \) of measurements determining the discrete prior.

The prior ‘graininess’ problem, which would show up as values of \( \chi^2 >> 1 \) for small \( N \) does not occur, because of the assumption that measurement uncertainty is large compared with population variability. This problem would occur if the height separation of cases constituting the prior was large compared with the measurement uncertainty, and none of the heights in the collection of cases constituting the prior was within measurement uncertainty of a height measurement. The Bayesian ‘universe of possibilities’ would then need to be expanded.
CONCLUSIONS

These types of calculations suggest that the discrete empirical Bayes prior does not significantly bias the posterior mean (Figure 5). The effect of uncertainty of the measurements determining the prior may be to cause some overestimation of the posterior standard deviation (Figure 6). The $\chi^2$ diagnostic is useful in detecting situations where the number of prior mixture components $N$ is too small, which might happen when the measurement uncertainty is small compared with the population variability.

In terms of real-world internal dosimetry, a crude version of discrete empirical Bayes has been in use at Los Alamos for many years now\(^{1,7}\) without evidence of serious difficulties. In this approach, for $^{239}$Pu inhalation intakes, the discrete set of biokinetic models in the biokinetic prior are ICRP-66 type M and S with particle sizes of 1, 5 and 10 $\mu$m AMAD (six models in all). Satisfactory values of $\chi^2$ are obtained for all cases in the Los Alamos Database (some 30,000 cases). However, for $^{238}$Pu, in order to have satisfactory values of $\chi^2$, the biokinetic prior needed to be expanded to include a peculiar ‘delayed-onset’ type of biokinetics\(^7,8\) associated with the ‘wing-9’ accident that occurred on 31 July 1971 involving high-fired ceramic material, as well as a variation of type-M behaviour observed in an accident that occurred on 16 March 2000. For the Mayak worker study\(^9\), a more rigorous application of the discrete empirical Bayes method is being used, with study cases provided by some of the large number of cases with autopsy tissue data.

ACKNOWLEDGEMENT

The author thanks David Pawel for the idea of using measurement of height as a heuristic example of the discrete empirical Bayes method.

FUNDING

This work is part of the United States-Russian Joint Coordinating Committee for Radiation Effects Research (JCCER) Project 2.5 and is funded under a Cooperative Agreement with the United States Department of Energy Office of International Health Programs (HS-14), Health Safety and Security Division (HSS). Funding to pay the Open Access publication charges for this article was provided by the United States Department of Energy contract for the management and operation of Los Alamos National Laboratory.

REFERENCES