Pure Appl. Chem., Vol. 75, Nos. 11–12, pp. 2151–2158, 2003.

Topic 3.13

Use of NOAEL, benchmark dose, and other models for human risk assessment of hormonally active substances*,†

R. Woodrow Setzer, Jr.[1,‡] and Carole A. Kimmel[2]

[1]USEPA, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Experimental Toxicology Division, Pharmacokinetics Branch, MD B143-01 109 TW Alexander Drive, Research Triangle Park, NC 27711, USA; [2]USEPA Office of Research and Development, National Center for Environmental Assessment (8623D), Ariel Rios Building, 1200 Pennsylvania Ave. NW, Washington, DC 20460, USA

Abstract: The benchmark dose (BMD) is the dose of a substance that is expected to result in a prespecified level of effect, the benchmark response level or BMR. It is a general approach to characterizing dose response, applicable to any toxicant and endpoint. A BMD is conceptually superior to a "no observed adverse effect level" (NOAEL) for this purpose because of being less determined by experimental design, because it is a precisely defined entity, and because its precision can be estimated. Since a BMD is a single number, just as an NOAEL, it is tempting to use the BMD as a straightforward replacement for the NOAEL in the assessment process for calculating allowable daily intakes. However, the level of toxic response at an NOAEL is unknown, while that at a BMD is well defined. Use of the BMD approach potentially adds consistency and objectivity to the process of deriving reference values (RfDs, RfCs, or ADIs) for setting regulatory levels. To take advantage of this, BMRs need to be selected in a consistent way across studies and endpoints. This paper discusses some issues affecting the selection of BMRs, and presents an example of a BMD calculated for the effects of peripubertal exposure to the fungicide vinclozolin.

INTRODUCTION

The benchmark dose (BMD) was originally proposed in 1984 [1] as an alternative to the NOAEL (no observed adverse effect level) and LOAEL (lowest observed adverse effect level) for setting regulatory levels such as reference doses (RfDs), reference concentrations (RfCs), and acceptable daily intakes (ADIs). The RfD, RfC, or ADI approach is used for agents whose dose-responses are thought to be very nonlinear or threshold-like. In this methodology, the regulatory level is derived by first determining a point of departure (POD) based on the dose response for the most sensitive endpoint(s) relevant to humans, then dividing it by a series of uncertainty factors (UFs) [2]. Typically, these uncertainty factors include a factor for extrapolation from animal to human data (UF_A), a factor to account for uncertainty

*Report from a SCOPE/IUPAC project: Implication of Endocrine Active Substances for Human and Wildlife (J. Miyamoto and J. Burger, editors). Other reports are published in this issue, *Pure Appl. Chem.* **75**, 1617–2615 (2003).

†The opinions expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

‡Corresponding author: E-mail: setzer.woodrow@epa.gov

about variability in the human population (UF_H), a factor for extrapolation from subchronic data to a chronic exposure scenario (UF_S), a factor for database deficiencies (UF_D), and a modifying factor (MF). Conventionally, NOAELs have been used as PODs, so when the lowest dose in the critical study is an LOAEL, an additional uncertainty factor has been used (UF_L). The default value for these uncertainty and modifying factors is 10, although an RfD or RfC with a total UF of >3000 is usually considered too uncertain to be reliable. Factors of 1, 3, or 10 are usually applied for the UFs depending on the available data, and chemical-specific pharmacokinetic and/or pharmacodynamic data can be used to adjust or replace these factors.

The BMD is used as an alternative to the NOAEL/LOAEL approach for a more quantitative way of deriving regulatory levels for health effects assumed to have a nonlinear (threshold-like) low dose–response relationship. Whereas NOAELs and LOAELs are discrete doses from a study, the BMD approach involves modeling the dose–response curve in the range of the observable data, and then using that model to interpolate an estimate of the dose that corresponds to a particular level of response, e.g., 5 or 10 % for quantal data, or some predefined change in response from controls for continuous data. A measure of uncertainty is also generally calculated, e.g., a confidence limit or Bayesian posterior [3], and the lower confidence limit on the dose used as the BMD is called the BMDL. The BMDL accounts for the uncertainty in the estimate of the dose response that is due to characteristics of the experimental design such as sample size. The BMDL is used as the basis for the point of departure (POD). In a recent health assessment done by the EPA for 1,3-butadiene [4], an additional factor for extrapolating from the BMDL to lower associated risk levels was applied (discussed further below).

Unlike NOAELs and LOAELs, BMDs are not constrained to be one of the experimental doses, and can thus be a more consistent basis for dose–response assessment. NOAELs do not correspond to a consistent response level and depend on sample size so that the NOAEL will be higher in studies with a smaller sample size, the opposite of what is desirable. In addition, NOAELs are usually associated with some definable level of risk, and are not threshold doses or "no effect levels". The slope of the dose–response curve is not usually considered in the NOAEL/LOAEL approach unless the slope is very steep or very shallow. If an NOAEL has not been defined in a particular study and only an LOAEL is available, an uncertainty factor is typically applied to account for the lack of an NOAEL (UF_L). The use of BMDLs as the POD has been the basis for several RfDs and RfCs in the IRIS database [12]. EPA's Draft Benchmark Dose Technical Guidance Document [6] outlines a number of considerations to be made in the derivation of BMDs and BMDLs.

The BMDL is also used as the basis for the POD for linear low-dose extrapolation, the dose–response assessment approach applied to most carcinogens [5]. In this case, once the POD is determined, risk is extrapolated linearly to a low dose corresponding to 10^{-5} to 10^{-6} risk.

The BMD can also be used to estimate relative potencies among different chemicals. Usually, in this case, the maximum likelihood estimate (BMD) rather than the lower confidence limit (BMDL) is used for the comparison.

The BMD approach can be used for dose–response modeling of all types of chemical and physical agents and associated endpoints, including endocrine active substances (EASs), regardless of the assumptions about low-dose linearity or nonlinearity. This is because dose–response modeling is done in the observable range and the BMD is typically related to a response rate near the lower end of the observable dose range. Whether the effect seen with EASs is a nonlinear or threshold-type response or is additive to background does not affect BMD calculation. Selection of the response level for deriving the BMD, i.e., the benchmark response, BMR, is the more difficult issue, especially for continuous endpoints.

ISSUES SURROUNDING SELECTION OF THE BMR

Type of data

Selection of the BMR depends on the kind of data being modeled, e.g., dichotomous (quantal) or continuous data. Other types of data are also encountered (e.g., categorical or graded responses), but are not dealt with specifically here. Such data are often converted to quantal data before modeling.

The approach to BMD development has been discussed most often in the literature for quantal data, primarily because a dichotomous response (i.e., whether a response is present or not) is somewhat easier to judge (e.g., tumor, malformation). BMRs have been expressed in terms of extra or additional risk, which are two ways of expressing the prevalence of adverse effects above background. EPA's draft BMD guidelines [6] recommend using extra risk as the more conservative approach. Thus, the BMD associated with an extra risk at a BMR above background is the dose where the following expression is true:

$$[P(d) - P(0)]/[1 - P(0)] \quad (1)$$

where P(d) is the risk at a dose = d and P(0) is the background risk at zero dose.

For continuous data, the BMR is expressed as a change in the mean from control values. The selection of the BMR is more difficult to determine for continuous data, because the goal is to base the BMR on a change that is biologically meaningful. However, for many endpoints, this degree of change has not been decided or agreed upon by relevant experts. The alternative, if no criteria have been developed for what degree of response is biologically meaningful, may be to use a change in mean response equal to one standard deviation of the control mean. Continuous data can be modeled as such or the data can be categorized (dichotomized) using the level of change considered meaningful or the 1 SD change, and modeled in the same way as for quantal data.

Selecting the benchmark response level

Once the dose response is established, the only factor that affects the magnitude of the benchmark dose is the selection of the BMR. How the BMR is selected can have a large effect on the way the resulting BMD is used in a subsequent risk analysis. For example, it is tempting to think of using the BMD like an NOAEL or LOAEL in a nonlinear risk assessment. With this interpretation, the BMD would be used along with the typical uncertainty and modifying factors appropriate for NOAELs or LOAELs. Several studies in the literature (e.g., [7–11]) have compared various approaches to setting the BMR by trying to ascertain the numerical relationship between BMDs calculated in different ways and NOAELs. However, this comparison was not meant to imply that BMDs should be direct substitutes for NOAELs. Rather, the comparison provided a reflection of the limit of detection and pointed out the differences in response rate at the NOAEL for different types of effects.

It is arguable that the approach to risk assessment based on NOAELs and LOAELs is weak because they are so strongly affected by the design of the bioassay from which they are generated. In addition, the level of effect actually present at the NOAEL and LOAEL is unknown and, therefore, they may not be as health-protective as desired. By contrast, the more objective BMD approach provides an opportunity to improve the consistency of risk assessments and their resulting health protectiveness.

One approach might be to select the BMR to reflect a constant level of toxicity, regardless of endpoint. This approach would certainly facilitate interpretation of the BMD, and would probably improve the consistency of risk assessments. However, there are problems with this proposal: for example, coming to common agreement on the amount of change in a continuous endpoint that is considered to be adverse; and, to some extent, the difficulty of equating changes in continuous endpoints with changes in the prevalence of discrete adverse effects. An alternative approach would be for experts to determine the degree of change that is considered biologically meaningful and adverse for each endpoint, thus removing the need to compare general levels of toxicity across endpoints. Taking the latter path would re-

quire a fair amount of effort to develop a consensus among regulatory toxicologists about adversity of responses for continuous data.

Given the difficulties that arise with the considered approaches to setting the BMR, it is tempting to simply fix upon a fixed default level of response, say, for example, 10 % increase in the prevalence of adverse effects, or a 1 SD change in the mean for continuous endpoints.

Use of BMDs in setting regulatory standards

Use of the BMD approach for nonlinear or threshold-type responses presents some challenges in terms of risk communication. This is because the BMD is associated with a particular level of response risk (5, 10 %), and the various uncertainty factors that are applied are not intended to reduce that risk, for the most part. On the other hand, BMD modeling can aid in determination of risk above the RfD and RfC values when exposures above those values occur.

Because UFs are applied multiplicatively they are acknowledged to overlap to some extent, so that the application of several UFs probably does effect some risk reduction. However, the intent of most of the UFs is not risk reduction. For example, the UF_A and UF_H deal, respectively, with the assumed differences in sensitivity between animals and humans, and the assumed variability in sensitivity among members of the human population. The UF_S and UF_D are intended to deal with various aspects of data deficiencies or limitations. Only the UF_L used for the LOAEL to NOAEL extrapolation is a true risk reduction factor. As indicated above, however, the NOAEL can be associated with a significant level of risk in animal bioassay studies (e.g., 5–40 %), and the UF_L may or may not adequately reduce the risk at the LOAEL to an acceptable level.

In a recent health assessment of 1,3-butadiene [4], a factor was applied to the BMDL for a quantal response in an attempt to reduce the risk associated with the POD and also to account for the slope at the BMD. This factor, which is a combination risk reduction factor and uncertainty factor, was termed the effect level extrapolation factor (ELF), and was applied to the POD. The UFs were then applied to this adjusted value. The ELF was determined as follows:

$$\text{ELF} = X \times (\text{slope from BMD}_x \text{ to } 0)/(\text{slope at the BMD}_x) \quad (2)$$

where X is the % incidence at the BMR, and BMD_x is the benchmark dose for x level of response. To account for uncertainties about the level of risk at the POD and to insure adequate reduction, the minimum factor applied is intended to be greater than 1 (usually a minimum of 3) up to X. The minimum factor is determined by consideration of the level of response at X, the weight of the evidence, and the endpoint(s) used to determine the POD. Thus, using this approach, RfDs and RfCs are more likely to represent a negligible level of risk. Probabilistic approaches to determining the range of uncertainty around the RfD and RfC may be useful in estimating the range of risk above the RfD or RfC when exposures occur above those levels [14–16]. However, these approaches have not been adopted on a regular basis in risk assessment as yet.

High-dose effects

However the BMR is selected, it is important that the effects selected be relevant to the human situation. This is obvious when considering on which endpoints to base a risk assessment, but is just as important when considering the BMR for a particular endpoint. It is not uncommon for the dominant mechanism of toxicity to change as the dose level increases. Unless the BMR is chosen in the range of toxicity that is relevant to the human exposure range, the BMD that results will not be an appropriate summary of the dose response.

Consideration of model uncertainty

In the paper in which he introduced the idea of benchmark dose, Crump [1] specified that the BMR should be selected to be in the range of the data. This minimizes the effect of model choice on the value of the BMD. Formally, one can divide the uncertainty of a BMD estimate into that due to the data itself, and that due to the uncertainty about the "true" model. The farther a BMR is selected from the responses present in the data, the more the overall uncertainty about the BMD value is due to this latter model uncertainty. Unfortunately, conventional statistical methods do not capture model uncertainty when they quantify the uncertainty of a parameter estimate. Thus, it is important to select a BMR where the model uncertainty contribution to the overall uncertainty is minimal.

EXAMPLE: PERIPUBERTAL EXPOSURE TO THE ANTIANDROGENIC FUNGICIDE VINCLOZOLIN

The fungicide vinclozolin and its two metabolites M1 and M2 are androgen antagonists. They produce adverse effects when administered during sexual differentiation in the fetus or around the time of puberty, and alter sexual function in adult male rats. The study from which the data for this example were taken [17] examined the effects of exposure to vinclozolin around puberty on the male reproductive tract and serum hormone levels. In this example, benchmark doses were calculated for age at preputial separation, epididymal weight, seminal vesicle weight, ventral prostate weight, and serum concentrations of testosterone and luteinizing hormone. The data were provided by Dr. Gray as group means, standard deviations, and sample sizes.

Computing a benchmark dose requires that: (i) a BMR be selected; (ii) one or several appropriate dose–response models be fit to the data; (iii) one model be identified based on an assessment of the quality of the model fit to the data; (iv) the best-fitting model be used to calculate the benchmark dose; and (v) confidence limits or credible limits be computed for the estimate. Serum LH will be used to illustrate these steps, which were followed for all six endpoints.

For this example, the BMD is the dose at which the mean of the response variable is expected to change by an amount equal to the standard deviation of the control group. This level of change very roughly corresponds to the increase in prevalence of extreme individual observations (that is, more extreme than a few percent) by about 10 % [18]. The maximum likelihood fit of a linear, quadratic, power, and Hill model [19], was determined for each endpoint using EPA's BMDS software (see next section for source). For each endpoint, two separate models for the within-dose-group variance were entertained: (i) the variance is the same for all dose groups; and (ii) the variance for a dose group is proportional to the mean raised to a power (e.g., if the estimate of the power is 2.0, the coefficient of variation is constant). Thus, a total of eight models were fit to each of the six endpoints. The best-fitting model of the set of eight, as determined by the sample size corrected version of the Akaike Information Coefficient (AIC_c) [20], was used to calculate the BMD and BMDL (Table 1).

For this example, we show some of the details of fitting the Hill model to the serum LH data. Table 2 presents the original means and standard deviations as well as those predicted by the model, and the scaled residuals: differences between observed and predicted means scaled by their predicted standard errors. These latter quantities are useful adjuncts for assessing model fit, since values with large absolute values (say, greater than 2) indicate points that are not well described by the model. Graphical evaluation is an essential element in assessing overall model adequacy: Figure 1 shows mean serum LH with 95 % confidence intervals and the fitted curve. Both the figure and the tabulated scaled residuals demonstrate that the model fits the data quite well, especially the data for the lowest dose. This confirms that the Hill model is a good choice for the LH data.

The BMDs and BMDLs for all six endpoints are graphed in Fig. 2. While the remaining endpoints seem to form a cluster, with similar BMDs that overlap each other's confidence limits, serum LH stands apart with a substantially lower BMD.

Table 1 AIC_c values for the eight models considered for each of the six endpoints. The minimum AIC_c value for each endpoint is underlined. For five endpoints, the power parameter in the power model was estimated to be 1.0, so it and the linear models resulted in the same BMD, BMDL, and AIC_c values.

Model[a]	Endpoint					
	Age at preputial separation	Epididymal weight	Serum LH	Seminal vesicle weight	Serum testosterone	Ventral prostate weight
lin-hom	<u>106.59</u>	<u>339.77</u>	38.49	<u>430.45</u>	79.51	<u>317.69</u>
lin-het	108.45	341.64	18.96	432.62	<u>71.36</u>	319.08
quad-hom	108.97	341.05	40.64	432.92	81.94	320.16
quad-het	111.01	343.62	17.72	435.24	73.74	321.64
power-hom	<u>106.59</u>	<u>339.77</u>	38.49	<u>430.45</u>	79.51	<u>317.69</u>
power-het	111.08	341.58	18.96	432.45	<u>71.36</u>	321.69
Hill-hom	109.74	340.86	40.58	435.54	81.94	320.16
Hill-het	111.81	343.05	<u>15.7</u>	437.85	73.71	324.64

[a]The word before the hyphen indicates the model for the mean (lin = linear, quad = quadratic, power = power, Hill = Hill model); the word after the hyphen indicates the variance model, either constant (homogeneous) or modeled as a power of the mean (heterogeneous).

Table 2 Mean serum LH concentrations (ng/ml) with their standard deviations and sample sizes, along with the values predicted by the fitted model and differences between the observed and expected mean LH levels scaled by the predicted standard error ("scaled residuals").

Dose	N	Mean		SD		Scaled Residuals
		Observed	Expected	Observed	Expected	
0	10	0.62	0.61	0.23	0.23	0.017
10	10	1.10	1.07	0.67	0.54	0.055
30	10	1.33	1.54	0.73	0.93	–0.234
100	10	2.30	2.06	1.59	1.45	0.163

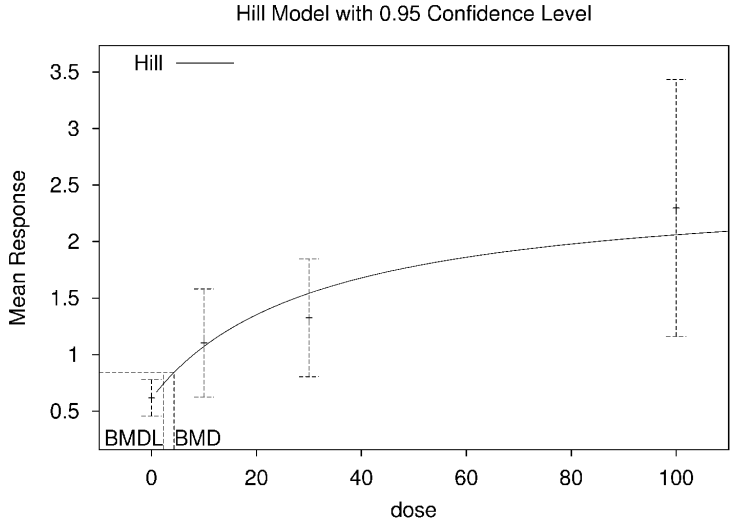

Fig. 1 Mean serum LH and 95 % confidence limits from [17]. The solid line is the fitted dose–response curve.

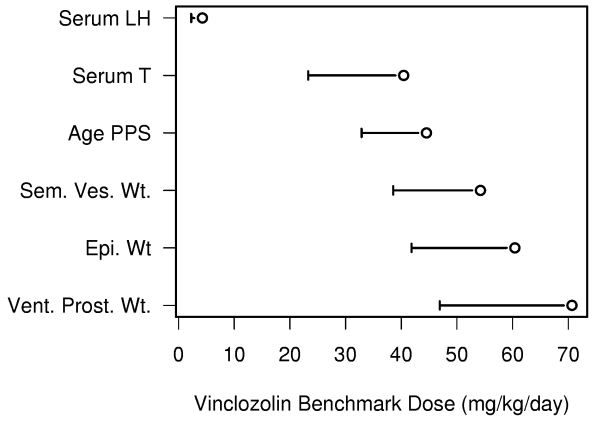

Fig. 2 BMDs and lower 95 % confidence limits for all six endpoints, based on the best-fitting model (determined by lowest AIC_c value) for each endpoint.

MODELING SOFTWARE FOR CALCULATING BENCHMARK DOSES

Many computer software packages allow nonlinear modeling of datasets, but most of these prove to be inadequate for benchmark dose modeling because they cannot calculate benchmark doses or their confidence limits. Currently (as of 1 October 2003), two packages have been created specifically for modeling toxicology data and calculating benchmark doses and their confidence limits:

BMDS: available for free download from the U.S. Environmental Protection Agency (<http://www.epa.gov/ncea>).

ToxTools: commercial software available from Cytel Software Corporation, Cambridge, MA (<http://www.cytel.com>).

In addition, BMD analysis can be carried out using general-purpose statistical software if it is flexible enough to allow the programming for calculation of the BMD and BMDL. This has the advantage that the analysis can be tailored to specific experimental designs and that there are no restrictions on the models that can be used. However, this approach requires substantially more statistical and programming skill than does using the special-purpose software.

RESEARCH PRIORITIES

There is much need for further research on the benchmark dose approach both in statistical methodology and in application to risk assessment. Methodologies need to be developed and applied to toxicology data for quantifying model uncertainty, which could allow extrapolation of the dose response to lower doses while tracking the uncertainty of doing so. Since health effects risk assessments are generally based on a review of effects on multiple endpoints based on multiple data sets, methods need to be developed for better modeling of multiple endpoints at a time, and for combining estimates across independent data sets.

On the applications side, more thought needs to be applied to the problem of extrapolating the BMD, usually derived from animal toxicology studies, to a safe human or ecological exposure level, especially for nonlinear effects.

REFERENCES

1. K. S. Crump. *Fundam. Appl. Toxicol.* **4**, 854–871 (1984).
2. D. G. Barnes and M. Dourson. *Regul. Toxicol. Pharmacol.* **8**, 471–486 (1988).
3. V. Hasselblad and A. M. Jarabek. In *Bayesian Biostatistics*, D. A. Berry and D. K. Stangl (Eds.), Marcel Dekker, New York (1995).
4. U.S. Environmental Protection Agency. National Center for Environmental Assessment-Washington Office, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC. EPA/600/P-98/001F, May (2002).
5. U.S. Environmental Protection Agency. National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC. NCEA-F-0644, July (1999). Available at <http://www.epa.gov/ncea/raf/cancer.htm>.
6. U.S. Environmental Protection Agency. National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC. EPA/630/R-00/001, October (2000). Available at <http://cfpub.epa.gov/ncea/>.
7. E. M. Faustman, B. C. Allen, R. J. Kavlock, C. A. Kimmel. *Fundam. Appl. Toxicol.* **23**, 478–486 (1994).
8. B. C. Allen, R. J. Kavlock, C. A. Kimmel, E. M. Faustman. *Fundam. Appl. Toxicol.* **23**, 487–495 (1994).
9. B. C. Allen, R. J. Kavlock, C. A. Kimmel, E. M. Faustman. *Fundam. Appl. Toxicol.* **23**, 496–509 (1994).
10. R. J. Kavlock, B. C. Allen, C. A. Kimmel, E. M. Faustman *Fundam. Appl. Toxicol.* **26**, 211–222 (1995).
11. J. R. Fowles, G. V. Alexeeff, D. Dodge. *Regul. Toxicol. Pharm.* **29**, 262–278 (1999).
12. U.S. Environmental Protection Agency. (2002). Online at <http://www.epa.gov/iris/index.html>.
13. U.S. Environmental Protection Agency. Vol. 1, EPA/600/R-99/001 (1997), and Vol. 2, EPA/600/R-98/155 (1998). Available at <http://www.epa.gov/nceawww1/colloquium.htm>.
14. S. J. S. Baird, J. T. Cohen, J. D. Graham, A. I. Shlyakhter, J. S. Evans. *Human Ecol. Risk Assess.* **2**, 79–102 (1996).
15. J. S. Evans, L. R. Rhomberg, P. L. Williams, A. M. Wilson, S. J. Baird. *Risk Anal*. **21**, 697–717 (2001).
16. J. C. Swartout, P. S. Price, M. L. Dourson, H. L. Carlson-Lynch, R. E. Keenan. *Risk Anal.* **18**, 271–82 (1998).
17. E. Monosson, W. R. Kelce, C. Lambright, J. Ostby, L. E. Gray, Jr. *Toxicol. Ind. Health* **15**, 65–79 (1999).
18. K. S. Crump. *Risk Anal.* **15**, 79–89 (1995).
19. U.S. Environmental Protection Agency. EPA/600/%-00/014F (2001).
20. K. P. Burnham and D. R. Anderson. *Model Selection and Inference*, p. 51, Springer Verlag, New York (1998).