International Journal of Occupational Medicine and Environmental Health, 2004; 17(1): 147—151

THE PRECAUTIONARY PRINCIPLE AND STATISTICAL APPROACHES TO UNCERTAINTY

NIELS KEIDING[1] and ESBEN BUDTZ-JØRGENSEN[1,2]

[1] Department of Biostatistics
Institute of Public Health
University of Copenhagen,
Copenhagen, Denmark
[2] Institute of Public Health
University of Southern Denmark
Odense, Denmark

Abstract. The central challenge from the Precautionary Principle to statistical methodology is to help delineate (preferably quantitatively) the possibility that some exposure is hazardous, even in cases where this is not established beyond reasonable doubt. The classical approach to hypothesis testing is unhelpful, because lack of significance can be due either to uninformative data or to genuine lack of effect (the Type II error problem). Its inversion, bioequivalence testing, might sometimes be a model for the Precautionary Principle in its ability to "prove the null hypothesis". Current procedures for setting safe exposure levels are essentially derived from these classical statistical ideas, and we outline how uncertainties in the exposure and response measurements affect the no observed adverse effect level, the Benchmark approach and the "Hockey Stick" model. A particular problem concerns model uncertainty: usually these procedures assume that the class of models describing dose/response is known with certainty; this assumption is, however, often violated, perhaps particularly often when epidemiological data form the source of the risk assessment, and regulatory authorities have occasionally resorted to some average based on competing models. The recent methodology of the Bayesian model averaging might be a systematic version of this, but is this an arena for the Precautionary Principle to come into play?

Key words:
Bayesian model averaging, Benchmark approach to safety standards in toxicology, Dose-response relationships, Environmental standards, Exposure measurement uncertainty, Popper falsification

INTRODUCTION

Barnett and O'Hagan [1] formulated the consequence of the Precautionary Principle for the use of scientific knowledge in setting environmental standards: "In general, we have seen that weak knowledge must lead to more stringent standards, on precautionary grounds. The value of science is to allow us, where appropriate, to relax the standards". In the present paper we analyze this aim in the context of several statistical approaches in environmetrics.

POPPER FALSIFICATION AND THE CLASSICAL STATISTICAL HYPOTHESIS TESTING THEORY

The central principle of Karl Popper's epistemology is that science progresses through the formulation of precise hypotheses which experimental data may then falsify. Following falsification the research must formulate new hypotheses which can then subsequently be tested versus new data from reality, etc. The significance test as formulated by Fisher [2] is very parallel to this concept. He stated: "A test of significance contains no criterion for "accepting" a hypothesis. Ac-

Received: January 19, 2004. Accepted: January 30, 2004.
Address reprint requests to Dr. N. Keiding, Department of Biostatistics, Institute of Public Health, University of Copenhagen, Blegdamsvej 3, DK-2200 Copenhagen N, Denmark (e-mail: nk@biostat.ku.dk).

cording to circumstances it may or may not influence its acceptability", and also emphasized that "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses".

Confronted with a putative environmental hazard, the Precautionary Principle does not find it satisfactory only to look for possible falsification of the claim that there is no hazard. This is because there are two very different reasons for the available data not containing evidence against this hypothesis: either there may indeed be no effect or an effect of a size below biological interest, or the data may be too noisy (often because there are too few of them) to enable any precise conclusion. The latter situation is nowadays termed "Type II error" in hypothesis testing theory and can be said to constitute the main problem in drawing precautionary conclusions from empirical evidence.

This dilemma has been noted in other areas of biostatistics, and a notable example is bioequivalence testing [3], which roughly speaking aims at proving that a new drug has the same effect as an old drug. A standard approach here is to twist classical hypothesis testing around: if θ is a measure of the difference between the new drug and the old drug, one postulates the null hypothesis

$$H_0: \theta < -\delta \text{ or } \theta > \delta$$

$$H_A: -\delta \leq \theta \leq \delta$$

If the null hypothesis H_0 is rejected, there is support for claiming bioequivalence; we falsify that the old drug and the new drug are different.

ARE THE EFFECTS OF STATISTICAL UNCERTAINTY ON CONVENTIONAL APPROACHES TO SETTING ENVIRONMENTAL STANDARDS IN ACCORDANCE WITH OR AGAINST THE PRECAUTIONARY PRINCIPLE?

We now turn more specifically to studying the effect of statistical uncertainty on three commonly used approaches to setting environmental standards.

No observed adverse effect level

The no observed adverse effect level (NOAEL) is a commonly used starting point for deriving reference doses from standard toxicological animal experiments. Doses are given at a relatively small number of discrete levels (including 0), and the NOAEL is defined as the highest dose that does not show increased risk over that of dose 0. This is usually interpreted in the standard hypothesis testing framework which implies that the more evidence (number of observations, precision of risk measure) the lower the NOAEL.

It is seen that the less evidence, the higher the standard, and we may therefore say that NOAEL is anti-precautionary.

The Benchmark approach

Crump [4] proposed the Benchmark approach to setting reference doses with one important motivation being the above mentioned anti-precautionary property of the NOAEL. The Benchmark approach defines the Benchmark dose (BMD) as the dose or exposure that corresponds to a specified increase (the Benchmark risk – BMR) in risk over the risk in an unexposed population. The BMD is estimated by fitting a dose-response model to the data. A statistical lower bound (BMDL) on the BMD then replaces the NOAEL in determining an exposure guideline [5]. It is easily seen that smaller studies lead to lower BMDL, so that in this sense the Benchmark approach is precautionary.

The Benchmark approach is amenable to application on epidemiological exposure-response data, and Budtz-Jørgensen et al. [6] gave a detailed survey of the literature and further discussion. In particular, these authors showed that the larger the variance around the dose-response curve, the larger the BMDL. This possibly somewhat counterintuitive property is a consequence of two opposing effects of increased random response variation: on one hand, the BMDL is lower as a result of increased estimation uncertainty, on the other hand, the response distribution becomes more dispersed, which will lead to a higher estimated BMD because the same increase in expected test performance corresponds to a smaller increase in the risk of an abnormal response. It is a mathematical fact that the latter effect will dominate the former, an anti-precautionary property.

It is useful in this connection to recall that Slob [7] emphasized that the variance around the dose-response curve should reflect the actual variation in the target population, which is often quite different from that underlying the calculations, particularly if these are based on toxicology experiments.

Budtz-Jørgensen et al. [8] have recently documented that measurement error in the exposure measurement leads to overestimation of BMD and BMDL, which in our parlance means that measurement error in the exposure variable makes the Benchmark approach anti-precautionary.

Hockey-stick models

A simple approach to setting standards is to postulate that the exposure-response relationship has the shape of a hockey-stick in the sense that it is zero up to a certain breakpoint and increases linearly thereafter (Fig. 1). To investigate the effect of the size of the experiment, the variation around the dose-response curve and measurement error in the exposure, we conducted a small simulation study reported in Table 1. It is seen that small experiments and large variance around the exposure-response relationship both lead to overestimation of the breakpoint. In our parlance these uncertainties make the hockey-stick models anti-precautionary. On the other hand, measurement error in the exposure decreases the estimate of the breakpoint, so that the hockey-stick model approach is precautionary in the face of this uncertainty.

Table 1. Simulation study*

Number of observations	Measurement error in dose		
	0%	0%	20%
	Residual variance in response		
	0.05	0.5	0.5
20	3.12	4.70	4.23
50	3.04	3.59	3.03
100	2.03	2.99	2.37

* Median estimate of the breakpoint in 250 Monte Carlo simulations of data sets from a hockey-stick model with a varying number of observations and different degrees of residual variation in the response and different degrees of measurement error in the exposure. The exposure measurement error variance is given in percent of the variance in the true exposure. In all simulations the true exposure is uniformly distributed from 0 to 10 while true breakpoint is 3.

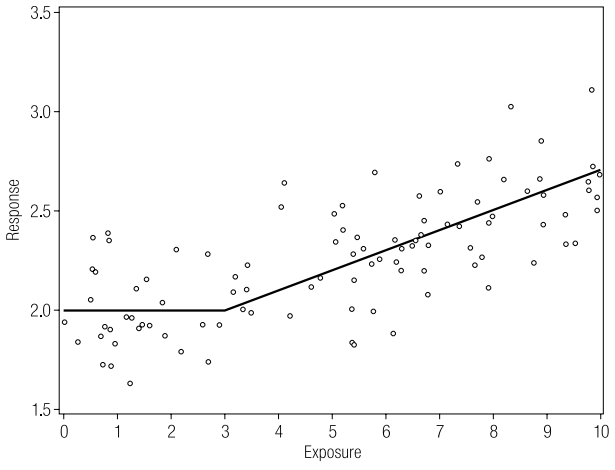

Fig. 1. Hockey-stick model. The exposure-response relation is zero up to a certain (unknown) breakpoint and increases linearly thereafter. In this case Response = 2 + 0.1 max(0, Exposure – 3) + error, where the number of observations is 100 and the variance of the error is 0.05.

MULTIPLICITY PROBLEMS

In some situations there are several manifest measures of an underlying latent exposure, and there may be several end-points. One such example is the study in the Faroe Islands on possible cognitive effects of prenatal exposure to methylmercury [9]. The mercury exposure of the fetuses was assessed both through the mercury concentration in the mother's hair at parturition and from the concentration of mercury in the cord blood. In addition, the amount of the mother's whale intake and fish intake during pregnancy were assessed through interview. The general end-point of "cognitive effects" was operationalized through an array of cognitive tests with no very specific prior hypothesis about the relative importance of these.

In such a situation, the classical statistical-epidemiological advice is to exercise considerable caution towards possible spurious effects of the inevitable "fishing expeditions" in the search for statistical significance in the data. Grandjean et al. [9] were well aware of this danger and took several kinds of precautions, notably by fixing a common set of confounders for all statistical tests, but the general problem remains.

One viewpoint is that there is essentially one latent active exposure variable, of which the various observed mercury concentrations are manifestations, and that there is one or a small number of cognitive response variables, collectively represented by the actual tests. Framed in this way

a structural equations approach is natural, and this was carried through in considerable detail by Budtz-Jørgensen et al. [10,11].

The Precautionary Principle may, however, encourage focus on the "most sensitive" end-point, which was indeed recommended in the practical implementation of the Faroe study by the National Research Council [12] in its report on toxicological effects of methylmercury. Crump [13] mentioned this issue in his overview of the application of the Benchmark approach to continuous response variables.

As documented by Budtz-Jørgensen et al. [10,11], the strength of the association between exposure and response was practically the same for the general effect as modeled by the structural equation approach as it was for the effect of cord blood mercury concentration on the most sensitive end-point. Implementation of the Benchmark approach in the structural equations context is under way.

MODEL UNCERTAINTY

Our final topic concerns the fact that there are in practice many choices to be made regarding the exact selection of statistical model. One issue regards selection of confounders, which we for the Faroe study briefly mentioned above. A systematic study was performed by Budtz-Jørgensen [14]. Standard covariate selection procedures produce a final model, and it is then conventional, albeit obviously untrue to quote standard errors as if this final model had been postulated without regard to the data. A detailed bootstrap analysis was able to take into account the selection uncertainty of various commonly recommended selection procedures and showed that for the case of the Faroe study the lowest honest standard error was obtained by keeping at least most covariates in the model. Indeed the compromise but systematic choice by Grandjean et al. [9] fared particularly well in this comparison.

Another matter regards the choice of exposure-response relationship. This is particularly tricky for epidemiological data where the zero-dose is usually poorly represented, so that although various models may fit almost equally well within the range of observed data, extrapolation to zero level can yield very different results. This uncertainty has led Crump [15] to recommend that attention be restricted to exposure-response relationships of the type ad^K for $K \geq 1$, based on a subject matter argument that values of $K < 1$ are unbiological. Budtz-Jørgensen et al. [6] documented a considerable uncertainty connected to using the Benchmark approach for the K-power models, a square root model and a logarithmic exposure-response model, and the National Academy of Sciences (NAS) report [12] concluded that a linear model was to be preferred on grounds similar to those advocated by Crump [15].

The biological arguments may not be definitively convincing in this case, and regulatory authorities have been tempted to produce some kind of average of the results of using various models [16]. From the viewpoint of the Precautionary Principle one could possibly argue that the exposure-response model yielding a reasonable fit and producing the lowest reference dose should be preferred, but if the environmental standard is to be determined completely or mainly on epidemiological evidence, this "maximin" approach may be too extreme from a general judgement.

The recent general statistical-methodological interest in the Bayesian Model Averaging [17] has been considered in the context of setting a standard for the content of arsenic in drinking water [18]. The general idea is to specify a set of suitable models, usually weighted equally *a priori*. From these posterior model probabilities are computed, reflecting the likelihood that a model holds, given the observed data. The results are then averaged with respect to these posterior probabilities, so that the better fitting models get upweighted at the expense of more poorly fitting models. This approach is heavily computer-intensive but feasible.

Our view regarding the suitability of this averaging approach for setting standards is cautious. If different models yield grossly varying safety limits, the most reasonable conclusion is not necessarily some kind of average, no matter how much its calculation may be based on general principles.

ACKNOWLEDGEMENTS

As is obvious from the above, this work owes much to our long and deep collaboration with P. Grandjean and P. Weihe. We are very grateful to Louise Ryan for giving

us access to her unpublished work on model averaging at short notice.

REFERENCES

1. Barnett V, O'Hagan A. *Setting Environmental Standards: the Statistical Approach to Handling Uncertainty and Variation.* London: Chapman & Hall; 1997.
2. Fisher RA. *Statistical Methods and Scientific Inference.* London: Oliver and Boyd; 1959.
3. Endrenyi L. Bioequivalence. In: Armitage P, Colton T, editors. *Encyclopedia of Biostatistics.* Vol. 1. Chichester: John Wiley & Sons; 1998. p. 368–72.
4. Crump K. *A new method for determining allowable daily intakes.* Fundam Appl Toxicol 1984; 4: 854–71.
5. Crump K. *Benchmark analysis.* In: El-Shaarawi AH, Piegorsch WW. Encyclopedia of Environmetrics. Chichester: Wiley; 2002. p. 163–70.
6. Budtz-Jørgensen E, Keiding N, Grandjean P. *Benchmark dose calculation from epidemiological data.* Biometrics 2001; 57: 698–706.
7. Slob W. *Deriving safe exposure levels for chemicals from animal studies using statistical methods. Recent developments.* In: Barnett V, Stein A, Turkman KF, editors. *International Forum, Statistical Aspects of Health and the Environment.* New York: John Wiley & Sons; 1999. p. 153–74.
8. Budtz-Jørgensen E, Keiding N, Grandjean P, et al. *Benchmark dose calculations adjusted for measurement error in the exposure variable.* Paper presented at the International Biometric Conference in Freiburg, 2002 July 21–26.
9. Grandjean P, Weihe P, White RF, Debes F, Araki S, Yokoyama K, et al. *Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury.* Neurotoxicol Teratol 1997; 19: 417–428.
10. Budtz-Jørgensen E, Keiding N, Grandjean P,Weihe P, White RF. *Statistical methods for the evaluation of health effects of prenatal mercury exposure.* Environmetrics 2003; 14: 105–20.
11. Budtz-Jørgensen E, Keiding N, Grandjean P, Weihe P. *Estimation of health effects of prenatal mercury exposure using structural equation models.* Environ Health 2002; 1: 2.
12. National Academy of Sciences (NAS). *Toxicological effects of methylmercury.* Washington, D.C: National Academy Press; 2000.
13. Crump K. *Critical issues in benchmark calculations from continuous data.* Crit Rev Toxicol 2002; 32: 133–53.
14. Budtz-Jørgensen E. *Statistical methods for the evaluation of health effects of prenatal mercury exposure* [Ph.D. Dissertation]. University of Copenhagen; 2001.
15. Crump K. *Calculations of benchmark doses from continuous data.* Risk Anal 1995; 15: 79–89.
16. Environmental Protection Agency (EPA). *Mercury Report to Congress.* Washington: Environmental Protection Agency, Office of Air Quality Planning and Standards; 1997.
17. Hoeting JA, Madigan D, Rafterry AE, et al. *Bayesian model averaging: A tutorial.* Statistical Science 1999; 14: 382–417.
18. Morales KH, Ibrahim JG, Chen CJ, Ryan LM. Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water [under revision for publication].