no toxicity and no tumors are produced. This has been most extensively investigated with respect to substances that produce urinary tract calculi (12). There are numerous such substances, including several which are essential for basic biological processes such as calcium, phosphate, cysteine, glycine, orotic acid, or oxalate (11,13). Numerous agricultural, commercial, and pharmaceutical chemicals also can produce calculi when there are high levels of exposure, such as melamine, terephthalic acid, nitrilotriacetate, sulfonamides, HIV protease inhibitors, and carbonic anhydrase inhibitors. Many of these have been known to produce calculi at high exposure levels, not only in experimental models but also in humans (11–13).

Urinary calculi can form directly from the administered chemical (or a metabolite) or can form from constituents normally present in the urine, such as calcium phosphate, calcium oxalate, or uric acid, secondary to alterations of normal physiological processes resulting in increased concentrations of these substances in the urine (13). Regardless of how the calculus is formed, it acts as an irritant, producing epithelial cytotoxicity, sometimes with full thickness ulceration of the epithelium and consequent hematuria. The amount of toxicity is dependant on numerous variables including the size of the calculi, their number, and the coarseness of the surface.

Sodium saccharin produces bladder tumors in rats when the administration begins at birth and continues for the life of the animal (14). In contrast, administration to mice or monkeys for their lifetime does not produce any effects on the lower urinary tract. In addition to being species specific, tumor formation is also a high dose phenomenon, requiring 25,000 ppm (2.5%) of the diet or higher. The mechanism of action involves dramatic changes in the urinary composition leading to the formation of calcium phosphate-containing precipitate (15). Calcium phosphate precipitate is cytotoxic to epithelial cells, including the urothelium. This leads to a cytotoxic effect on the bladder epithelium with consequent regenerative proliferation and ultimately tumors. Like the situation with calculi, this is a threshold phenomenon based on the solubility of calcium phosphate.

Numerous changes in the urine composition must occur for the calcium phosphate precipitate to form, and if any of these parameters are not affected appropriately, the precipitate does not form. Thus, administering saccharin as the acid rather than as the sodium salt produces an acidic urine which inhibits the formation of the calcium phosphate precipitate (16). Thus, there is no precipitate formation, no toxicity, and ultimately no tumors. In the mouse, the urinary concentration of calcium and phosphate is 10–20 times lower than that of the rat, and is not high enough for precipitate to form, again, a clear indicator of the threshold phenomenon that is involved (15). It turns out that in humans, like monkeys, the urine does not contain adequate amounts of protein nor is the osmolality sufficiently high for the precipitate to occur (15,17). Thus, sodium saccharin is both a species specific and a high dose (threshold) chemical carcinogen.

Inorganic arsenic is also carcinogenic for humans, and the organic arsenical, dimethylarsinic acid (DMA) is carcinogenic in the rat (18). It also illustrates the complex nature of the various phenomena included in the term genotoxicity.

Genotoxicity can involve direct interaction of the chemical with DNA, referred to as DNA reactivity, such as in the case of the aromatic amines (3). Arsenicals do not bind to DNA (19).

Although arsenic is not DNA reactive, numerous *in vitro* studies have demonstrated that it does produce various forms of genotoxicity, resulting primarily from inhibition of DNA repair, oxidative damage, or binding to tubulin, which can indirectly affect DNA (20). Although DNA reactivity can theoretically be nonthreshold (see below), these other forms of genotoxicity all involve threshold phenomena (21,22). Thus, it is critical in discussing genotoxicity to specify what type of phenomenon is involved with respect to a given chemical. All except DNA reactivity are clearly threshold phenomena.

In the case of arsenicals, many of the studies showing various types of genotoxicity actually involve concentrations *in vitro* that are higher than concentrations required to kill the cells (18–20). Likewise, *in vivo*, the dose required to demonstrate genotoxicity is usually higher than the dose necessary to produce a proliferative or tumorigenic response. Thus, it is unlikely that arsenicals are carcinogenic by a genotoxic mode of action.

Instead, it is more likely that arsenicals are carcinogenic to the bladder and other tissues by a mode of action involving cytotoxicity and regenerative proliferation (18,23). Arsenicals are metabolized by a sequence of reductions of the pentavalent to trivalent form followed by oxidative methylation of the trivalent species. The pentavalent forms of arsenic are generally quite inactive with respect to toxicity, whereas the trivalent forms are extremely reactive, frequently being lethal to cells *in vitro* at concentrations less than 1 μM. Dimethylarsinous acid (DMA^{III}) and monomethylarsonous acid (MMA^{III}) are particularly cytotoxic to cells. Administration of arsenite, arsenate, or DMA^{V} to rats at high doses produces a high concentration of trivalent arsenicals, particularly DMA^{III}, in the urine which is cytotoxic to the urothelium leading to regenerative hyperplasia. In mice, arsenate and arsenite are able to produce a similar cytotoxicity and regeneration response (23,24) whereas DMA^{V} does not (18), likely be-

cause of its more limited metabolism in the mouse compared to the rat. Similar to other substances which are carcinogenic by a mode of action involving cytotoxicity and regenerative proliferation, it is likely that arsenic carcinogenesis also involves a threshold, despite having some genotoxic but not DNA reactive properties.

DNA reactive carcinogens remain a unique group of chemicals. Their dose response has generated considerable controversy both with respect to genotoxicity and carcinogenicity for several decades. An experiment referred to as the megamouse experiment performed in the 1970's at the National Center for Toxicological Research was designed to try to address this issue (6,25). The reason for it being called the megamouse experiment was that more than 24,000 mice were utilized, with several hundred per dose group so that the level of detection for a carcinogenic response was 1% rather than the usual approximately 10% when the standard 50 or 60 animals are used per group (25). The carcinogen 2-acetylaminofluorene (AAF) was administered in the diet and sacrifices were performed at 18, 24 and 33 months. The doses used in the experiment, 0, 30, 45, 60, 75, 100, and 150 ppm were considerably lower than in previous experiments with AAF.

Quite unexpectedly, the dose response for the two target tissues of AAF in mice, the liver and urinary bladder, was completely different (6,25). The dose response in the liver was nearly linear, whereas the dose response in the bladder had an apparent threshold of approximately 45 to 60 ppm, with a statistically significant incidence detected at 60 ppm but not at 45 ppm. The question remained however, did this represent a true threshold or was the level of detection, even at 1%, inadequate to detect the low incidence of tumors that might occur at the lower doses in the urinary bladder. Based strictly on the dose response curve for tumors, it would appear that this was a threshold phenomenon, since the dose response was similar to that seen with clearly threshold-type carcinogens, such as calculi-forming chemicals, sodium saccharin, or arsenic. Although there is an apparent threshold, this can only be ascertained by evaluating the detailed mechanism that is involved in the carcinogenic response.

Since AAF is a DNA reactive carcinogen, it forms DNA adducts. In an experiment by Beland *et al.* (26), they were able to demonstrate that the dose response for DNA adducts was linear, both for the liver and the urinary bladder, even down to doses that were considerably lower than those used in the megamouse experiment. The DNA adducts were determined at a steady state level after one month of administration.

For the liver, at the concentrations used in this experiment, AAF produces DNA adducts in the normal hepatocytes, but is metabolized to a much lesser extent and forms adducts to a lesser extent in the foci (which represent the intermediate cell population, one step toward the formation of a hepatocellular tumor) (6). At these low concentrations, there is no evident cytotoxicity or other evidence of increased cell proliferation. Thus, the only effect is on DNA damage in the first step in the carcinogenic process, increasing the probability of a critical mistake occurring in the DNA with each replication normally occurring in the hepatocyte population. The background hepatocellular proliferation rate is approximately 50 to 100 days.

In contrast, DNA adducts form in all cells of the bladder since the reactive intermediate is generated in the liver and excreted in the urine (6). Thus, the probability of critical mistakes occurring during DNA replication is increased not only in the normal bladder epithelial cells, but also in cells all along the process in the development of cancer. The number of cells, however, in contrast to the liver, is much fewer, with only approximately 25,000 stem cells in the normal bladder epithelium and several million in the liver. Even with the number of animals in this experiment, the detection rate for increased tumor incidence is still only 1 % above background. At doses of less than 60 ppm in the diet, there are likely to be an insufficient number of tumors generated in this population to develop a statistically significant incidence. However, at doses of 60 ppm, there is an increase in the rate of cell proliferation and an increase in cell number (hyperplasia). This greatly potentiates the effects of the DNA reactivity, since it increases the number of replicating cells that are present as well as having an increase in the probability of critical DNA damage each time the DNA replicates.

Modeling of such processes estimates the tumor incidence to be expected taking into account both DNA adduct formation and DNA replication (6). In normal bladder epithelial cells, the rate of proliferation is extremely low, similar to the liver. At the dose of 60 ppm and above, there is not only an increase in the proliferation rate, but with hyperplasia, there is an increase in the number of cells. Thus, the number of DNA replications is greatly increased. In these modeling efforts, it can be shown that at doses of 60 ppm and above, a statistically significant incidence of tumors above 1% would occur. The apparent threshold can be shown to be not a true threshold, since there are DNA adducts formed and presumably DNA damage at much lower doses. However, utilizing these modeling systems, it can be shown that the expected tumor incidence at doses of 45 ppm and below would yield incidences well below the 1% that is the detection limit for this experiment.

Thus, although the shape of the tumor dose response curve of AAF is similar to that for calculus-forming rodent bladder carcinogens, based on mechanism, one can not conclude definitely that there is a true threshold.

The key events in the process of AAF-induced car-

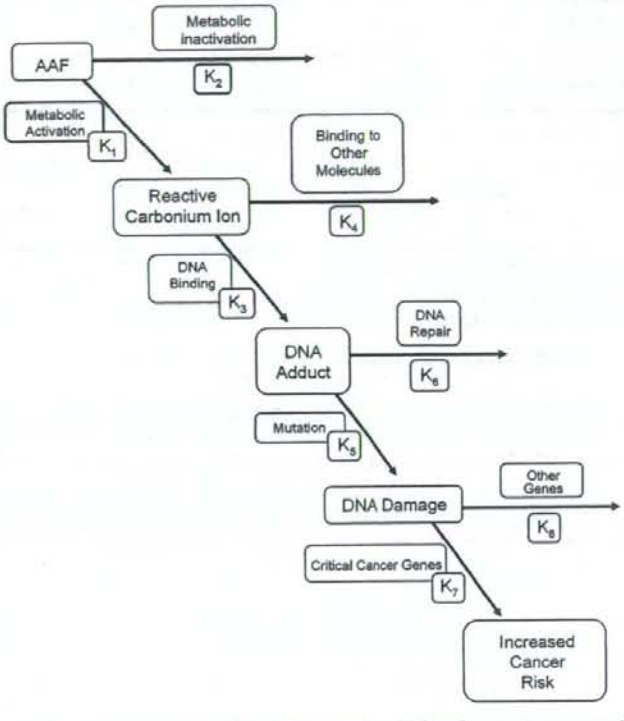

Fig. 1. Competing activating and inactivating processes in AAF carcinogenesis, with imaginary rate constants for each step. Only if all of the activating rate constants (odd numbered constants in the diagram) are zero at some dose can there be a true threshold.

cinogenesis include metabolic activation to a reactive intermediate which leads to the formation of DNA adducts, some of which are not repaired and lead to a mutation in critical cells, ultimately leading to an increased risk of cancer development. However, for each of these key events, there are competing events that would lead to an inactivation of the process (Fig. 1). Thus, not only is there metabolic activation, but there are numerous enzymes involved in metabolic inactivation of AAF. Once the reactive intermediate has formed, it can bind to a variety of chemicals, not only DNA, but protein, RNA, or even water. Binding to any of the substances that are not DNA obviously would not increase the risk of development of cancer. Furthermore, once DNA adducts form, many of them are repaired, again leading to inactivation of the carcinogenic effects of AAF. For those that are not repaired, permanent DNA damage will only occur if the cell is replicating at the time the adduct is present. Furthermore, the DNA adduct could form in one of the genes critical to the development of cancer, or form in one of the many other nucleotides available in the genome. These latter DNA adducts would not increase the risk of cancer development. For cancer to develop, all of the DNA alterations that are required for the development of cancer must occur in a single cell and that cell must be in the stem cell population. Furthermore, there are a variety of cellular repair processes once the critical DNA damage has occurred. These competing events are illustrated in Fig. 1. Imaginary rate constants are given for each of these competing processes, with odd numbered rate constants for the activating processes and even numbered rate constants given for the inactivating processes. It is only when all of the activating rate constants (odd numbered) are zero can a true threshold be possible. It is my impression that one or more of these rate constants at low concentrations can be zero, but we do not yet have the technology with appropriate detection limits to be able to prove the possibility of thresholds for DNA reactive carcinogens.

There are several terms which are used in a confusing fashion and sometimes interchangeably that must be distinguished carefully when addressing the issue of threshold. Most importantly is the issue of threshold vs. level of detection. Also, many individuals have used the terms threshold and non-linearity synonymously, whereas in fact they are distinguishable as illustrated in the example of 2-AAF above. A critical distinction is a true threshold vs. a practical threshold. A true threshold, although likely for DNA reactive carcinogens, has yet to be proven as far as I am aware.

However, we regularly deal with the concept and validity of practical thresholds. For example, numerous natural DNA reactive carcinogens are present in our diet, such as aflatoxin. Because of dramatic developments in technology, we can measure incredibly small amounts of aflatoxin so that it can be detected in virtually all peanut products as well as in many other grain products. Whether or not there is a true threshold in the carcinogenic effects of aflatoxin is unknown. However, a safe level can be estimated for practical purposes so that we can consume peanuts and other products which contain miniscule amounts of the chemical. Such estimates are based on an extrapolation to low doses based on animal and/or human investigations, with an extrapolation estimate of overall risk of 1 in 100,000 or 1 in 1,000,000 individuals. This is then set as a safe level for regulatory purposes, and serves as a practical threshold for carcinogenicity.

In summary, thresholds for genotoxicity are known to occur for genotoxic mechanisms not involving direct DNA reactivity. For DNA reactivity, the issue remains unresolved. Similarly, for carcinogenesis, non-DNA reactive genotoxic and completely non-genotoxic chemical carcinogens clearly have thresholds. For DNA reactive carcinogens, again, this remains unresolved.

Acknowledgments: I am grateful to Connie Winters for her assistance in preparation of this manuscript and to Lora Arnold for her helpful critique.

References

1 Cohen SM, Shirai T, Steineck G. Epidemiology and etiology of premalignant and malignant urothelial changes. Scand J Urol Nephrol. 2000; 205: 105–15.
2 Clayson DB, Cooper EH. Cancer of the urinary tract. In: Klein G and Weinhouse S, editors. Advances in cancer research. New York: Academic Press, Inc.; 1970. p. 271–381.
3 Miller EC, Miller JA. Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules. Cancer. 1981; 47: 2327–45.
4 Greenfield RE, Ellwein LB, Cohen SM. A general probabilistic model of carcinogenesis: Analysis of experimental urinary bladder cancer. Carcinogenesis. 1984; 5: 437–45.
5 Cohen SM, Ellwein LB. Genetic errors, cell proliferation, and carcinogenesis. Cancer Res. 1991; 51: 6493–505.
6 Cohen SM, Ellwein LB. Proliferative and genotoxic cellular effects in 2-acetylaminofluorene bladder and liver carcinogenesis: Biological modeling of the ED_{01} study. Toxicol Appl Pharmacol. 1990; 104: 79–93.
7 Cohen SM. Cell proliferation and carcinogenesis. Drug Metabolism Rev. 1998; 30: 339–57.
8 Cohen SM. Urinary bladder carcinogenesis. Toxicol Pathol. 1998; 26: 121–7.
9 Meek ME , Bucher JR, Cohen SM, DellarcoV, Hill RN, Lehman-McKeeman LD, Longfellow DG, Pastoor T, Seed J, Patton DE. A framework for human relevance analysis of information on carcinogenic modes of action. Crit Rev Toxicol. 2003; 33: 591–653.
10 Andersen ME, Meek E, Boorman GA, Brusick DJ, Cohen SM, Dragan YP, Frederick CB, Goodman JI, Hard GC, O'Flaherty EJ, Robinson DE. Lessons learned in applying the U.S. EPA's proposed cancer guidelines to specific compounds. Toxicol Sci. 2000; 53: 159–72.
11 Clayson DB, Fishbein L, Cohen SM. The effect of stones and other physical factors on the induction of rodent bladder cancer. Fd Chem Toxicol. 1995; 33: 771–84.
12 IARC Working Group. Consensus Report. International Agency for Research on Cancer, IARC Scientific Publications. 1999; 147: 1–32.
13 Cohen SM, Johansson SL, Arnold LL, Lawson TA. Urinary tract calculi and thresholds in carcinogenesis. Food Chem Toxicol. 2002; 40: 793–9.
14 Ellwein LB, Cohen SM. The health risks of saccharin revisited. Crit Rev Toxicol. 1990; 20: 311–26.
15 Cohen SM. Calcium phosphate-containing urinary precipitate in rat urinary bladder carcinogenesis. International Agency for Research on Cancer, IARC Scientific Publications. 1999; 147: 175–89.
16 Cohen SM, Ellwein LB, Okamura T, Masui T, Johansson SL, Smith RA, Wehner JM, Khachab M, Chappel CI, Schoenig GP, Emerson JL, Garland EM. Comparative bladder tumor promoting activity of sodium saccharin, sodium ascorbate and related acids and calcium salts in rats. Cancer Res. 1991; 51: 1766–77.
17 Cohen SM. The role of urinary physiology and chemistry in bladder carcinogenesis. Food Chem Toxicol. 1995; 33: 715–30.
18 Cohen SM, Arnold LL, Eldan M, Schoen AS, Beck BD. Methylated arsenicals: The implications of metabolism and carcinogenicity studies in rodents to human risk assessment. Crit Rev Toxicol. 2006; 36: 99–133.
19 Nesnow S, Roop BC, Lambert G, Kadiiska M, Mason RP, Cullen WR, Mass MJ. DNA damage induced by methylated trivalent arsenicals is mediated by reactive oxygen species. Chem Res Toxicol. 2002; 15: 1627–34.
20 Kligerman AD, Tennant AH. Insights into the carcinogenic mode of action of arsenic. Toxicol Appl Pharmacol. 2007; 222: 281–8.
21 Kirkland D, Pfuhler S, Tweats D, Aardema M, Corvi R, Darroudi F, Elhajouji A, Glatt H, Hastwell P, Hayashi M, Kasper P, Kirchner S, Lynch A, Marzin D, Maurici D, Meunier JR, Muller L, Nohynek G, Parry J, Parry E, Thybaud V, Tice R, van Benthem J, Vanparys P, White P. How to reduce false positive results when undertaking in vitro genotoxicity testing and thus avoid unnecessary follow-up animal tests: Report of an ECVAM Workshop. Mutat Res. 2007; 628: 31–55.
22 Zeiger E. History and rationale of genetic toxicity testing: an impersonal, and sometimes personal, view. Environ Mol Mutagen. 2004; 44: 363–71.
23 Cohen SM, Ohnishi T, Arnold LL, Le XC. Arsenic-induced bladder cancer in an animal model. Toxicol Appl Pharmacol. 2007; 222: 258–63.
24 Suzuki S, Arnold LL, Ohnishi T, Cohen SM. Effect of in-

organic arsenic on the rat and mouse urinary bladder. Toxicol Sci. 106: 350–63.

25 Littlefield NA, Farmer JH, Gaylor DW, Sheldon WG. Effects of dose and time in a long-term, low-dose carcinogenic study. J Environ Pathol Toxicol. 1979; 3: 17–34.

26 Beland FA, Fullerton NF, Kinouchi T, Smith BA, Poirier MC. DNA adduct formation in relation to tumorigenesis in mice chronically fed 2-acetylaminofluorene. Prog Clin Biol Res. 1990; 331: 121–9.

Genes and Environment, Vol. 30, No. 4 pp. 139–149 (2008)

Mini-review

Experimental Design and Statistical Analysis of Studies to Demonstrate a Threshold in Genetic Toxicology: A Mini-review[1]

David P. Lovell[2]

Postgraduate Medical School, University of Surrey, Surrey, UK

(Received September 25, 2008; Revised October 25, 2008; Accepted October 26, 2008)

A mechanistic understanding of genotoxicity is important for the risk assessment of the exposure of human populations to chemicals. The nature of the dose response relationship at low doses is valuable information in the evaluation of the biological importance of such exposures. A range of mathematical and statistical approaches have been used to try to characterize responses at these low doses. Methods include mathematical models which do or do not include thresholds and statistical methods which try to identify No-observable effect levels (NOELs). It is important to appreciate that determination of an NOEL is not evidence for a threshold. There is an increasing appreciation of the potential to identify 'pragmatic' thresholds using experimental systems with a range of biomarkers. The accurate characterization and estimation of these dose-response relationships requires careful experimental design which can improve the accuracy of the estimates of the response while avoiding the introduction of artifactual effects. Statistical approaches such as Design of Experiment (DoE) methodology, which builds on the traditional factorial design, can provide efficient approaches for the description and estimation of dose-response relationships of both individual and combinations of agents. Estimation approaches such as the benchmark dose methodology and the concept of thresholds of toxicological concern provide practical methods for addressing the threshold problem.

Key words: statistics, experimental design, threshold, genotoxicity

Introduction

The objective of this paper is to provide an overview of the statistical and experimental design issues involved in the design and interpretation of studies to identify thresholds associated with exposure to genotoxic agents.

It has become an axiom that genotoxic chemicals induce DNA damage at any level of exposure and do not have a threshold in their dose-response relationships. Madle *et al.* (1), for instance, stated that "...it is generally agreed that there are no thresholds for genotoxic effects of chemicals, i.e., that there are no doses without genotoxic effects." This concept is the basis of risk assessment strategies for chemicals with genotoxic potential. These consider that genotoxic carcinogens do not have a threshold while those that act by non-genotoxic mechanisms may have a threshold (the default assumption in other areas of toxicology). Chemicals with thresholds are regulated via limit values such as ADI (Allowable Daily Intake) or TDIs (Tolerable Daily Intake) while non-thresholded chemicals are regulated using concepts such as ALARP (As Low As Reasonably Practical) or ALARA (As Low As Reasonably Achievable). (For recent reviews, see (2,3)).

Genetic damage can be gene mutation, chromosome damage (clastogenicity) and chromosome loss (aneuogenicity) and is detected by batteries of short-term mutagenicity tests. Increasingly, some aneuogens, based upon their mechanism of action (MOA), are considered to have thresholds but that, in the absence of evidence to the contrary, gene mutagens do not. Chemicals, however, may have a number of MOAs including both direct and indirect action on the DNA.

Many of the ideas relating to low dose modelling such as the Linear Non-Threshold (LNT) dose-response model derive from radiation biology (4,5). The concepts remain contentious and there continues to be an active debate in the field (5,6). Gene mutations are assumed to have linear kinetics because they arise from single (one-hit) events with the dose-response relationship consequently being linear at low doses. Chromosomal damage may result from two or more hits (such as a chromosomal break followed by a rearrangement) with a linear-quadratic (non-linear or curved) dose-response relationship. It is assumed that there is a non-zero, although small, chance that a single 'hit' of radiation

[1]Presented at the International Symposium on *Genotoxic and Carcinogenic Thresholds*, Tokyo, Japan, July 22-23, 2008.

[2]Correspondence to: David P Lovell , Postgraduate Medical School, University of Surrey, Daphne Jackson Road, Manor Park, Guildford, Surrey, GU2 7WG UK. Tel: +44-1483-688609, Fax: +44-1483-688501, E-mail: d.lovell@surrey.ac.uk

will damage genetic material which then results in a mutagenic event leading ultimately to a cancer. This translates into the concept of 'no safe dose of radiation' and subsequently into the 'one molecule can cause cancer' concept applied to chemicals (7).

In practice, descriptions of what occurs at very low doses whether in terms of molecular or statistical models are theoretical and may only partially reflect what is actually happening. Even at very low doses millions of molecules of a compound may be involved.

Genomic DNA is exposed to continual 'attack' by endogenous mutagens (such as reactive oxygen species) which may result in some 'spontaneous' or 'background' genetic damage. This can provide a reference level for the assessment of the potential 'added' risk that might arise from a low level of exposure leading, perhaps, to a '*de minimis*' dose based upon some change compared with the spontaneous level of damage.

Different types of genetic damage may show different types of dose-response relationships. Biomarkers of exposure like DNA adducts appear to show linear responses in low-dose studies using accelerator mass spectrometry and other mass spectrometry methods while biomarkers of effect such as gene mutations may show non-linear responses because of defence mechanisms such as DNA error repair and detoxification. The dose-response relationships for adducts and gene mutations are not parallel and the relationship between the two markers may be complex (8,9).

Methods exist for comparing the induction of chromosome loss and non-disjunction in combination studies which allow the exploration of effects at low doses (10). Experimental evidence of thresholds has been provided for a number of chemicals including spindle poisons and toposiomerase II inhibitors (11–13). Thresholds may exist for aneuogens because they act through the disruption to the protein structure making up the spindle such as by binding to tubulin. Thresholds may arise because a critical number of target sites must be affected before the effect occurs and that there is some redundancy in the target. The evidence that alkylating genotoxins may have thresholds has recently been reviewed (14).

Various terms have been suggested to qualify the term threshold. Kirsch-Volders *et al.* (15) describes "absolute", "real or biological", "apparent" and "statistical" thresholds. Lovell (16) has argued for 'practical' or 'pragmatic' thresholds while Hengstler *et al.* refer to "perfect" and "practical" thresholds (17). Jenkins *et al.* (14) noted that terms like absolute, biological, apparent, acceptable, statistical, NOEL, real, alleged, were used to qualify the term.

The ICPEMC (International Commission for Protection against Environmental Mutagens and Carcinogens) stated that "A threshold dose-response relationship is one in which a range of sub-critical doses is incapable of producing the specified response; as dose increases, the minimal dose that can elicit the response is the threshold dose" (18). ECETOC (The European Centre for Ecotoxicology and Toxicology of Chemicals) defined an 'absolute' threshold as "... a concentration below which a cell would not 'notice' the presence of the chemical. In other words, the chemical is present but does not interact with the cellular target" (15).

Thresholds are well known in pharmacology for many receptor-activation-dependent processes and with non-carcinogenic endpoints because of detoxification and error repair mechanisms. There are also proponents of the concept of U-or J-shaped curves where high doses are toxic while low doses are protective. This concept of biphasic responses or hormesis has been reviewed by Calabrese and co-workers (19,20). Examples include adaptive responses where it is thought that a low priming dose may reduce the effect of a second higher dose. Effects such as 'bystander effects', where a biological effect occurs not in the cell that has been 'hit' but in one in close proximity, and 'genomic instability' have been suggested as explanations at the mechanistic/cellular level of possible threshold effects (reviewed by Preston (21)).

Pragmatic attempts to move forward from the no safe dose/no threshold default assumption for genotoxic chemicals have included the use of the concepts such as de minimis (from the phrases *de minimis non curat praetor* or *de minimis non curat lex* taken to mean that the law is not interested in trivial matters) and the virtually safe dose (VSD) often associated with it and the concept of the threshold of toxicological concern (see later).

Linear and Non-linear

The terms linear and non-linear can cause some confusion. In the context of graphical presentation of dose-response data linear is equated with a 'straight line' relationship where a change is directly proportional to the exposure. The slope, in effect, is the regression coefficient, which represents the change in the dependent variable for each unit change in the independent variable. Non-linear is often used to refer to a curved relationship where the relationship between exposure and effect is more complex. A simple definition of non-linear is where the effect is disproportionate to the cause. Non-linear relationships can take many forms. It is important to appreciate that the term non-linear is not synonymous with a threshold relationship.

Dose-response relationships can be referred to as linear, sub-linear (or convex as the relationship curves below the linear) or supra-linear (concave as it curves above the linear) (Fig. 1). The sigmoid curve is a combination of both. Supra-linear may indicate decreased toxic effects or saturation at higher doses; sub-linear

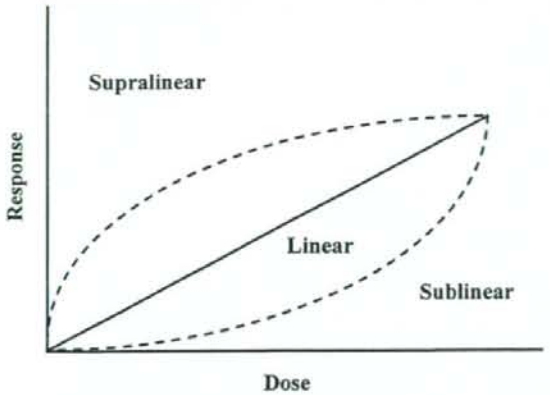

Fig. 1. Illustration of linear, sub-linear and supra-linear curves.

may indicate repair or deactivation at low doses. Lutz *et al.* proposed statistical tests for sub-linearity at low doses (22).

Linear and some non-linear relationships are monotonic. A monotonic relationship is where the responses at higher dose levels are always equal to or greater than the responses at lower dose levels. A monotonic relationship can be a prerequisite for some model fitting programs/software.

In the context of statistical models there is a distinction between linear and non-linear regressions. Non-linear refers to a mathematical model where the parameters values need to be fitted by iteration. Non-linear regression is an interactive approach to fitting a curve through data. It uses iterative (trial and error) methods which require an initial value, a statistical approach (e.g. maximum likelihood) and some measure of when the method (e.g. goodness of fit test) has converged (or has failed to converge). Such methods are now tractable given the increase in computing power. Linear regression is the statistical methodology for fitting the best fitting line through a set of points defined by the intercept and slope and does not involve iteration. The standard approach minimizes the sum of squares of the distance between the points and the line. The least squares approach generalizes to multiple regression where there are multiple independent variables fitted to the dependent variable.

A statistical linear model is one that can be expressed in a linear form. An example is the General Linear Model (GLM) which links a series of apparently unconnected statistical methods. For example, the common two-sample t-test is a special case of the analysis of variance methodology (which links to ancova and MANOVA) which, in turn, is related to linear and multiple regression methods through the GLM. The GLM is a special case of generalized linear models (GZM) (23) which extends the type of data, that can be analysed using linear models by using link functions (forms of transformations) which are related to the underlying distribution of the data.

Transformations

A transformation is a process for preparing data for analysis. Examples include the use of the log dose or converting a response to the change from control or baseline. Transformations aim to simplify the mathematics, ensure that the underlying assumptions are met, allow linear modelling, stabilize the variances and linearize the relationship for presentation of data as straight lines which is more convenient for interpretation.

The apparent shape of the dose response relationship depends upon how the data are graphed. Visual inspection (or 'ocular regression' (24)) of the dose-response relationship can be misleading and identifying whether a threshold is present cannot be determined just by graphing the data. It is, therefore, important to check how the data are actually presented. Beware of 'optical illusions' as the pattern will change using raw dose, log dose and extended log dose. Note particularly any discontinuities in the dose axis. Many graphs produced by Excel have equal spaced points on the X axis independent of the actual dose. This can create a graph which is based on neither the original nor the log transformed dose metric.

Transformation (by changing the scale) of either the X or Y axes can change a straight line into a curved line or vice versa. Lutz *et al.* (25) noted that "Logarithmic representation of the dose axis transforms a straight line into a sublinear (up-bent) curve, which can be misinterpreted to indicate a threshold." Fig. 2 shows examples of a dose-response relationship plotted when one or both scales are converted to log scales: raw/raw plot, log/raw plot, raw/log plot and log/log plot.

Problems can arise over the presentation of zero on the graph as the log of 0 cannot be calculated. Statistical analysis of trend tests using log doses need a substitute value for the zero dose level to carry out an analysis. The software package Graph Pad, for instance, suggests using a value about 2 log units below the lowest "real" X value on a log X scale. The presentation of data using dose metrics and its effect on low dose extrapolation has resulted in considerable debate between Waddell and others (26–36).

Slob discussed the concept of a practical threshold and argued that an attempt to identify one is not possible (37). He stressed that trying to show experimental evidence of a threshold was impossible. He showed how the GST-P positive foci data of Wanibuchi *et al.* (38) on MeIQx could be presented in 4 difference ways. One relationship was thresholded, one linear, another supralinear and the fourth sublinear. He noted that us-

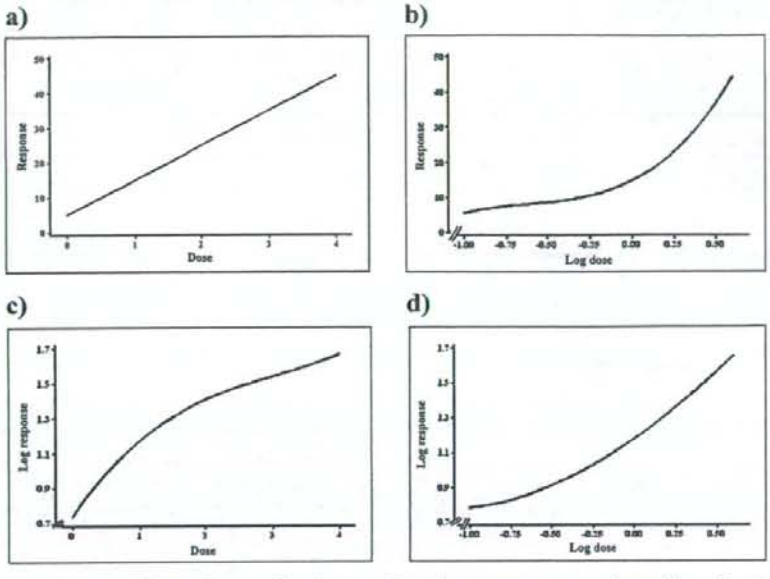

Fig. 2. Illustration of data plotted on different dose metrics: a) untransformed responses v. untransformed doses b) untransformed responses v. log 10 transformed doses c) log 10 transformed responses v. untransformed doses d) log 10 transformed responses v. log 10 transformed doses. Hypothetical data to illustrate how transforming the axis of graphs can change the shape of the dose response relationship.

ing the alternative presentations provided no evidence of a threshold and that the interpretation of the data would be changed depending upon which was used. His preferred or 'best way' of presentation of the data was on a double log scale plot arguing that the log scale for the response may be more appropriate because it represented a relative change compared with an absolute change. He stressed the benefit of the benchmark dose approach (see later) compared to the pair-wise statistical testing NOAEL approach.

Curve Fitting and Modelling

A range of methods have been developed for fitting curves to data. Complex curve fits can be achieved by approaches including polynomials, splines and Lowess regression curves (39,40,41). An important point about approaches such as the use of polynomials is that the predicted response values can become negative outside the experimental data range producing unrealistic extrapolations. A number of statistical packages including GraphPad and Statistica have the capability to fit many different curves to data.

The threshold or 'Hockey stick' model is an example of a piece-wise linear regression (42). Specific models to detect discontinuities, inhomogeneities, change points or inflections in a relationship such as a time-series are of interest in areas such as climate forecasting and financial markets. Identifying such discontinuities is not simple.

Much of modern statistical work is centred on modelling. This is a continuous process of building, refining, testing and modifying the models. Box (43) reviewing the contributions of RA Fisher, the geneticist and statistician, to modelling as part of a description of the philosophy for statistical modelling used the phrase "All models are wrong", which he subsequently modified to "All models are wrong (but) some are useful"(44).

The objective of model building is to explain a relationship and then use this to make predictions. Approaches such as multiple regression include Forward, Reverse, Stepwise and Best-subset approaches (45). Using the Forward approach parameters are added to the model until there is no improvement in the model fit; the Reverse approach starts with a full model and terms are dropped until the fit is reduced. The Stepwise is a combination of the two approaches. These different methods can result in different models. In general, the more parameters the better the fit, leading eventually (if enough parameter can be fitted) to a perfect fit. One of the potential problems of model fitting is 'over fitting' where the model provides a good fit to the observed data but provides poor predictivity (45).

Many of the model fitting methods have a goodness of fit measure such as a P value which indicates the degree by which the predicted values differ from the observed values. Small P values usually indicate a poor fit to the data. P value of less than 0.10, rather than the

usual 0.05 or 0.01, have been suggested in the benchmark dose modelling software as the critical value for defining a poor fit. P values >0.10 do not imply a correct model only that the results from the model are consistent with the observed data. Many models may give adequate fits to the observed data. The goodness of fit tests are not suitable for distinguishing between different models and the size of the P value provides no evidence for distinguishing between models.

Formal testing is only possible if the competing models form a family of models (nested or hierarchical models). In the case of models where the same method of fitting is used, Likelihood ratio tests and Akaike's information criterion (AIC) can be used to test whether the addition or subtraction of a parameter improves the fit of the model. The tests compensate for the increased fit associated with increasing terms in the model. Although useful for help in model selection the methods should be used with care. If the models are not from the same family, such as, for instance, the probit and logistic models, the goodness of fit tests cannot be used.

A key aspect of model fitting is the need to check the assumptions underlying the model such as independence of the data, homogeneity of variances and normality of residuals are met using a set of 'diagnostic' tools. Graphical methods should be used in conjunction with the statistical tests. Judgement and experience is needed in the modelling process: for example, the choice of whether to include or exclude data from the high doses in the model or to drop results of a dose to achieve monotonicity. The top dose may be influential in determining the best fitting model but if the fit to the data at top dose is less good than at lower doses then the model may be inappropriate. Ideally there should be results for doses near the likely benchmark dose to provide more accurate estimates. Model uncertainly should be addressed. In practice a number of models with a similar number of parameters can give satisfactory fits to the same data set. In the case of the benchmark dose approach the range of estimates obtained from all the acceptable models can be presented.

Modellers often work to the principle of parsimony or Occam's razor: "the simpler the explanation, the better". The overall objective should be to find a model which provides a satisfactory description of the dose-response data using the minimum number of parameters. So if the linear model is not significantly worse that the threshold or quadratic one then it could be argued that the simpler more parsimonious linear model should be used. However, the key message is that ultimately any model must not just provide a good fit but also good prediction when tested with new data.

In the context of prediction a distinction is sometimes made between the terms interpolation and extrapolation. Interpolation refers to the estimation of effects within the range of doses measured (between a pair of known data points); extrapolation relates to estimation outside the range. In general, it is usually assumed that interpolation should be reasonably reliable but that extrapolation (which also has a wider qualitative interpretation) outside the observed range needs to be carried out with care as there is a risk of serious error. An example is how the predictions from polynomials can become negative outside the observed range and differ greatly from the 'target curve'. There is a debate on whether estimation between a low and the negative control or spontaneous level is interpolation or extrapolation.

Thresholds and the NOEL/NOAEL

The most important point in this section is that identifying a No Observable Effect Level (NOEL) or No Observable Adverse Effect Level (NOAEL) using a statistical approach does not mean that a threshold exists.

The NOEL (and NOAEL) is widely used in the assessment of non-cancer endpoints. However, the NOEL should not be equated with a threshold dose below which effects do not occur. The NOEL/NOAEL depends upon specific characteristics of the experiment: sample size, statistical test, dose spacing. The limitations of the NOEL are well known (46) and has, in part, resulted in interest in the development of the benchmark dose methodology (below).

Determination of the NOEL is often based upon a hypothesis test. A statistical test based upon a test of a null hypothesis is able to show a positive effect (a statistically significant increase) but not to show a negative effect i.e. that the null hypothesis is true. Failure to find statistical significance shows only that there is not sufficient evidence to reject the null hypothesis. It is a serious error to equate a non-significant result with a lack of effect.

The danger of equating statistical significance with a threshold is increased by the use of multiple comparison methods (such as the Bonferroni correction) in a dose-response study to adjust the significance levels reported. These methods are aimed at avoiding type I errors (falsely declaring a result as significant when it is not) when making many statistical comparisons. In practice, the approach 'dampens' down the significance levels and effectively increases the dose at which the NOEL is identified. The choice of a critical value associated with a P value of 0.05 is also arbitrary and probability levels of 0.01 or lower are sometimes used as the cut-offs to designate effects as significant. Combining multiple comparison levels with more stringent significance levels can change what is considered the NOEL appreciably especially if there are many dose groups in a dose-response experiment as the Bonferroni correction will become more severe as it is dependent upon the number of groups in an experiment.

Background/Control Incidence

The size of the spontaneous or background level of a biomarker has implications for the statistical power of the study (especially for qualitative endpoints). Effects can be either absolute or relative changes. The absolute size of a fold-change, obviously, depends upon the control incidence: with control units of 2 or 20 units, the fold change is 2 or 20 units respectively.

The argument has been made that low-dose linearity will result if just some of the cancers caused by a chemical and those in the control group were indistinguishable because they were a consequence of the chemical induced damage being produced by the same mechanism as that which produced the spontaneous or 'background' cancers (47). The argument has been generalized further by Crawford and Wilson (48) who suggested that low dose linearity is generalizable to a much wider set of non-cancer outcomes.

Another argument has been that while there may be a threshold at the individual animal level there will be a distribution of tolerances so that some individuals may not have a threshold. Lutz (49), for instance, argued that genetic heterogeneity in a population means that at least some members of a population will not have a threshold for a toxic agent. He argued that this breaks down the distinction between both carcinogenic and non-carcinogenic agents and between genotoxic and non-genotoxic carcinogens.

Experimental and Statistical Issues in Designs at Low Doses and Curve Fitting Design Issues

The standard regulatory tests are designed for hazard identification where the objective is to produce a qualitative–positive or negative–result rather than risk estimation. Standard statistical methods such as pair-wise comparisons or trend tests such as for linear regression or the Cochran-Armitage trend test are often used. Tests of a linear trend in a dose-response in a design are statistically more powerful than pair-wise comparisons because of the natural or inherent ordering imposed on it by the experimenter but need a more specific null hypothesis.

The results will depend upon the choice of the dose metric. This may be the applied (mg/kg), the log dose (which needs a non-zero dose for the control group to circumvent the problem of taking the logarithm of zero) or the target dose (i.e. the dose delivered to the target tissue, related to the pharmacokinetics of the substance). Some care may be needed in the interpretation if the linear and/or the quadratic components in the analysis are significant.

An alternative approach is to estimate the size of an effect. As discussed above while experiments cannot be designed to show that a threshold exists (i.e. to prove the null hypothesis of no difference) the alternative statistical approach of estimating the size of effect can produce a bound on the size of effect that could be detected by a particular design. The size of effect could then be put into perspective with both the background level and the degree of variability around this measure. In this context some concept of effects which may be real but are small enough to treat as unimportant can lead to a *de minimis* approach. The benchmark dose approach (discussed later) is one such approach.

Identifying confidence intervals of a specific width around the dose-response relationship requires an appropriate and careful design. Biomarkers such as genotoxic endpoints are potentially sensitive to producing artefacts such as can arise through non-randomization. Therefore, if the objective is to investigate low level effects then careful experimental design is critical.

The traditional experimental methods of randomization, replication, blocking and local control (originally developed by Fisher) remain important in controlling variability. Fisher's concept of the factorial design which has extended into Design of Experiment (DOE) methodology and optimal design is a highly efficient and cost-effective approach to the investigation of complex problems and is appreciably more efficient than the use of the traditional One Factor at a Time (OFAT) approach (50).

The more experimental/dose groups the more precise will be the description of the dose-response relationship with multiple dose levels in the area of interest (16).

Using 6 or more dose groups, with perhaps fewer than the typical 5–6 animals per group, will enhance the precision with which the overall dose-response curve can be estimated (51).

An assumption of most standard statistical tests: t-test, anova, chi-square, Fisher's exact test is that the experimental units are independent. The concept of the experimental unit is fundamental to the statistical analysis of designed experiments. It is the unit in the experiment that is randomly assigned to the treatment. Mis-specification of the experimental unit can lead to serious misinterpretation of the statistical analysis.

It is important to appreciate that while the individual cell may be the smallest unit which can be measured, cells from the same animal or culture are not randomly assigned to the treatment but rather receive the same treatment and are likely to show some degree of correlations in their responses. A failure to appreciate this can lead to errors in analysis and interpretation. In particular failure to take to take into account hidden level of variation can lead to serious false positives.

The design of studies for characterising effects (estimation) is different from that for hazard identification with more doses required and more resources in the area of interest but there are implications in attempting to power the study to show no or small genotoxic effects.

Table 1. Sample size per group associated with 80% and 90% power for detecting an effect in standard deviation (SD) units in a two sided test at P = 0.05

SD units	Power	
	80%	90%
0.0625	**4020***	5381
0.1	1571	2103
0.125	**1006**	1346
0.2	394	527
0.25	**253**	338
0.3	176	235
0.4	100	133
0.5	**64**	86
0.6	45	60
0.8	26	34
1	**17**	23
1.25	12	15
1.5	9	11
2	**6**	7
2.5	4	5
3	4	4
4	3	3

*Bold numbers illustrate the approximate four-fold increase in sample size with each halving of the effect size.

Table 1 shows the size of effect that can be detected for a given power. A rule of thumb is that for 80% power for a 2-sided test at P = 0.05 then a 4-fold increase in sizes is needed for each halving of effect size. For an effect equal to one standard deviation (SD) the group sizes are approximately 16 for 80% power, for an 0.5 SD unit difference the group sizes need to be about 64. Historical data on the inter-experimental unit variability can be used to provide estimates of the size of a standard deviation unit.

The level of effect detectable in an experiment is illustrated using GST-P positive foci data from Table 3 in Murai *et al.* (52). The negative control mean is 22.9 foci with approximate inter-individual within group standard deviations of about 5 foci. Experiments with sample sizes of 50 would have 80% power to detect a difference from the control level of about 0.57 SD units or about 2.8 foci (in a pair-wise comparison using a two-sided test at P = 0.05).

If the determination of an effect were based solely on whether there is a significant difference between two concentrations then there might be different NOAELs for each biomarker because each biomarker would have different sensitivities or 'resolutions' based upon their 'intrinsic' variability. Studies of an equivalent size with a qualitative endpoint such as whether a tumour is present or absent will have lower power: a fact long appreciated in the interpretation of cancer bioassay data.

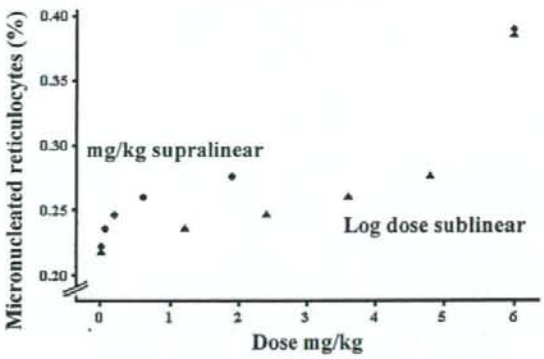

Fig. 3. Plot of mean of groups from Table II of Asano *et al.* (54) plotted against mg/kg dose (●) and against dose on log 10 dose scale (▲) (As doses were logarithmically spaced on the mg/kg dose scale the zero dose was represented by a dose level an equivalent spacing below the lowest dose on the log dose scale and the log dose scale aligned with the mg/kg dose scale.)

Example of a Test System to Explore Thresholds

The development of flow cytometry for the analysis of micronuclei (both *in vitro* and *in vivo*) is a potentially powerful tool for exploring thresholds. Automated scoring procedures allow previously unattainable numbers of cells to be readily scored (53). It is now possible to 'score' over a million cells per animal. Such an approach can appear to be very powerful. Tourus *et al.* (53) suggested that by scoring up to 3 million cells from a single sample that it is possible to detect a difference between 0.10% and 0.11% (based upon simulation experiments using samples with 'spiked malarial micronuclei') using Fisher's exact test. (It is important to note, though, that Fisher's exact test assumes independence of the measures.) Such large potential sample sizes means that, while measuring more cells from an experiment makes the estimate for that specific animal or culture more precise, apparently significant differences between dose levels could be a consequence of artifacts.

Asano *et al.* (54) in an *in vivo* experiment analyzed micronuclei for 1-β-D-arabinofuranosylcytosine using flow cytometry. Analysis was carried out on 1M, 200K, 20K and 2000 cells per animal using automatic and manual scoring. Scores for individual animals were reported by Asano *et al.* The results are presented as parallel curves using log log scales and show apparently increased variability between animals at 2K cells. Graphical presentation of the data shows the different shaped dose-response curves (Fig. 3). Using the untransformed dose the response is supra-linear while the curve is sub-linear against the logarithm of the dose illustrating that care is needed with visual inspection of

Table 2. Analysis of variance table of orthogonal breakdown of individual animal scores from raw data provided in Table II of Asano *et al.* (54)

a) Using original dose metric (mg/kg)

Source	df	MS	F	P
Between groups	5	0.019	5.58	0.002
Linear	1	0.091	27.36	<0.001
Quadratic	1	0.000	0.01	0.97
Cubic	1	0.001	0.42	0.91
Deviations	2	0.000	0.06	0.95
Within groups	24	0.003		
Total	29	0.173		

b) Using log dose

Source	df	MS	F	P
Between groups	5	0.019	5.58	0.002
Linear	1	0.068	20.37	<0.001
Quadratic	1	0.016	4.91	0.036
Cubic	1	0.007	2.12	0.16
Deviations	2	0.002	0.24	0.79
Within groups	24	0.003		
Total	29	0.173		

Table 3. Summary of comparisons between negative control and dose groups from Asano *et al.* (54)

Dose (mg/kg BW)	Cells				Mice
	2K	20K	200K	1M	1M
0.06					
0.19			*	*	
0.60			**	**	
1.89		**	**	**	
6.00		**	**	**	**

*P < 0.05
**P < 0.01

Data from Table I of Asano *et al.* (54) of frequencies of micronucleated reticulocytes of various doses of 1-β-D-arabinofuranosylcytosine. Statistical comparisons using the cell as the statistical unit carried out using Fisher's exact test, comparison using the animal as the statistical unit carried out using Student's t-test.

graphs.

In both cases in the analysis there is a significant linear component to the dose-response with also a significant quadratic component for the log dose but this is not significant for the mg/kg dose (Table 2). Statistical analysis of the pair-wise comparisons using the cell as the unit showed significant effects at lower doses but with the animals as the experimental units only at the top dose because of appreciable between animal variability (using hierarchical/nested analyses) (Table 3). Abramsson-Zetterberg (55) also used flow cytometry in two experiments with a large number of dose groups of acrylamide which showed a linear dose response with no evidence of a threshold.

Benchmark Dose Approach

Benchmark dose methodology is an alternative approach to the NOEL/NOAEL (56). The Benchmark Dose (BMD) is the dose associated with a pre-defined biologically significant/important difference, the Benchmark Response (BMR), in an endpoint of interest compared with the negative control response. The BMR could be a 10% change in average adult body weight or a doubling of a liver enzyme level. The BMD is derived by fitting a dose-response model to the experimental data. The BMD related to a change in response equal to one standard deviation from the control mean is also estimated to help with the comparisons.

The BMDL is the lower one-sided 95% confidence limit or bound on the BMD. Using the BMDL is a way of expressing that there is 95% confidence that the true effect at this dose would be less than the effect associated with the BMDL. The BMDs and BMDLs from the models are compared for similarity and consistency. The chosen BMDL is used as the Reference Point (RP) from which low dose extrapolation begins in order to find a reference dose (RfD) such as an ADI.

The NOEL/NOAEL is one example of an RP or Point of Departure (PoD) used to identify an ADI using safety or uncertainly factors (SF/UF).A number of organization such as EFSA and JECFA have proposed the use of BMD as the RP for the calculation of the Margin of Exposure (MOE) of genotoxic carcinogens.

Thresholds of Toxicological Concern

A pragmatic approach to the problem of no safe level of a genotoxic agent is the concept of the threshold of toxicological concern (TTC). The TTC is an intake by humans that would be associated with a high probability of negligible risk. The TTC was originally developed as a method for regulating food contact materials as a concentration level below which there would be no appreciable risk to human health irrespective of whether or not there are chemical-specific toxicity data (reviewed by Kroes *et al.*, (57)). Based upon studies on a database of carcinogenic chemicals a daily exposure of less that 1.5 μg/day was considered a virtually safe dose in that in a worst case scenario this corresponded to a 10^{-6} lifetime risk. To accommodate possible higher risk chemicals an exposure level of 0.15 μg/day was proposed (58).

The underlying approach is based upon a statistical analysis of a database of the carcinogenic potencies of a large number of chemicals calculated on the basis of a linear extrapolation from the dose (TD50) estimated to give a 50% tumour incidence. The probability distribution of the potencies was used to identify a threshold concentration which had a high probability of negligible risk (59, 60). The threshold is thus a pragmatic concen-

tration to help with regulation rather than indicating absolute certainty of no risk below the concentration.

The European Medicines Agency (EMEA) (61) has proposed that concentrations of less than 1.5 μg/day (corresponding to a 10^{-5} lifetime risk) would be acceptable for genotoxic impurities and contaminants in pharmaceuticals where there is a risk:benefit consideration. Humphrey has argued for a more flexible approach than a standard figure (62).

Conclusions

A wide range of curves and mathematical models can be fitted to experimental data but cannot prove the existence of an absolute threshold in a dose-response relationship. Equating a threshold with the identification of a NOEL/NOAEL based upon statistical significance using a hypothesis testing has serious limitations. Statistical methods based upon the estimation of the size of a response together with the associated confidence intervals can provide estimates of doses where there is high confidence of negligible risk. Approaches based upon the benchmark dose methodology and threshold of toxicological concern have the potential to be used in this way. Future advances in the field may include the development of more sophisticated mathematical models which include consideration of DNA repair and metabolic detoxification (63) and the use of multivariate methods in association with -omics technologies to investigate the pattern of responses in low dose studies.

References

1 Madle S, von der Hude W, Broschinski L, Jänig GR. Threshold effects in genetic toxicity: perspective of chemicals regulation in Germany, Mutat Res. 2000; 464: 117–21.

2 Barlow S, Renwick AG, Kleiner J, Bridges JW, Busk L, Dybing E, Edler L, Eisenbrand G, Fink-Gremmels J, Knaap A, Kroes R, Liem D, Müller DJG, Page S, Rolland V, Schlatter J, Tritscher A, Tueting W, Würtzen G. Risk assessment of substances that are both genotoxic and carcinogenic: Report of an International Conference organized by EFSA and WHO with support of ILSI Europe. Fd Chem Toxicol. 2006; 44: 1636–50.

3 O'Brien J, Renwick AG, Constable A, Dybing E, Muller DJ, Schlatter J, Slob W, Tueting W, van Benthem J, Williams GM, Wolfreys A. Approaches to the risk assessment of genotoxic carcinogens in food: A critical appraisal. Fd Chem Toxicol. 2006; 44: 1613–35.

4 Coggle JE. Biological effects of radiation. London: Taylor and Francis; 1983.

5 ICRP ICRP Publication 60. 1990 Recommendations of the International Commission on Radiological Protection, Annals of the ICRP 21. 1991; 201.

6 Wakeford R. The risk to health from exposure to low levels of ionising radiation, Annals of the ICRP 35. 2005; v-vii.

7 Efron E. The apocalyptics. New York: Simon and Schuster; 1984.

8 Perera EP. The significance of DNA and protein adducts in human biomonitoring studies. Mutat Res. 1988; 205: 255–69.

9 Zito R. Low doses and thresholds in genotoxicity: from theories to experiments. J Exp Clin Cancer Res. 2001; 20: 315–25.

10 Parry JM, and E.M. Parry EM. The use of the in vitro micronucleus assay to detect and assess the aneugenic activity of chemicals. Mutat Res. 2006; 607: 5–8.

11 Elhajouji A, Van Hummelen P, Kirsch-Volders M. Indications for a threshold of chemically-induced aneuploidy in vitro in human lymphocytes. Environ Mol Mutagen. 1995; 26: 292–304.

12 Elhajouji A, Tibaldi F, Kirsch-Volders M. Indication for thresholds of chromosome non-disjunction versus chromosome lagging induced by spindle inhibitors in vitro in human lymphocytes. Mutagenesis. 1997; 12: 133–40.

13 Bentley KS, Kirkland D, Murphy M, Marshall R. Evaluation of thresholds for benomyl- and carbendazim-induced aneuploidy in cultured human lymphocytes using fluorescence in situ hybridization. Mutat Res. 2000; 464: 41–51.

14 Jenkins GJS, Doak SH, Johnson GE, Quick E, Waters EM, Parry JM. Do dose response thresholds exist for genotoxic alkylating agents? Mutagenesis 2005; 20: 389–98.

15 Kirsch-Volders M, Aardema M Elhajouji A. Concepts of threshold in mutagenesis and carcinogenesis. Mutat Res. 2000; 464: 3–11.

16 Lovell DP. Dose-response and threshold-mediated mechanisms in mutagenesis: statistical models and study design. Mutat Res. 2000; 464: 87–95.

17 Hengstler JG, Bogdanffy MS, Bolt MH, Oesch F. Challenging dogma: thresholds for genotoxic carcinogens? The case of vinyl acetate. Annu Rev Pharmacol Toxicol. 2003; 43: 485–520.

18 Ehling DAUH, Cerutti PA, Friedman J, Greim H, Kolbye AC Jr, Mendelsohn ML. Review of the evidence for the presence or absence of thresholds in the induction of genetic effects by genotoxic chemicals. Mutat Res. 1983; 123: 281–341.

19 Calabrese EJ, Baldwin LA. Hormesis: U-shaped dose responses and their centrality in toxicology. Trends Pharmacol Sci. 2001; 22: 285–91.

20 Calabrese EJ, Blain R. The occurrence of hormetic dose responses in the toxicological literature, the hormesis database: an overview. Toxicol Appl Pharmacol. 2005; 202: 289–301.

21 Preston RJ. Bystander effects, genomic instability, adaptive response, and cancer risk assessment for radiation and chemical exposures. Toxicol Applied Pharmacol. 2005; 207: 550–6.

22 Lutz RW, Stahel WA, Lutz WK. Statistical procedures to test for linearity and estimate threshold doses for tumor induction with nonlinear dose-response relationships in bioassays for carcinogenicity. Regul Toxicol Pharmacol. 2002; 36: 331–7.

23 Nelder J, McCullagh P. Generalized linear models, Lon-

don: Chapman and Hall; 1989.
24 Greenland S. Dose-response and trend analysis in epidemiology. Epidemiology. 1995; 6: 356–65.
25 Lutz WK, Gaylor DW, Conolly RB, Lutz RW. Nonlinearity and thresholds in dose-response relationships for carcinogenicity due to sampling variation, logarithmic dose scaling, or small differences in individual susceptibility. Toxicol Appl Pharmacol. 2005; 207: 565–9.
26 Waddell WJ. Thresholds of carcinogenicity of flavors. Toxicol Sci. 2002; 68: 275–9.
27 Waddel WJ. Comparison of human exposures to selected chemicals with thresholds from NTP carcinogenicity studies in rodents. Hum Exp Toxicol. 2003; 22: 501–6.
28 Waddell WJ. Thresholds in chemical carcinogenesis: what are animal experiments telling us? Toxicol Pathol. 2003; 31: 260–2.
29 Waddell WJ. Threshold for carcinogenicity of N-nitrosodiethylamine for esophageal tumors in rats. Fd Chem Toxicol. 2003; 41: 739–41.
30 Waddell WJ. Thresholds of carcinogenicity in the ED01 study. Toxicol Sci. 2003; 72: 158–63.
31 Waddell WJ. Critique of dose response in carcinogenesis. Hum Exp Toxicol. 2006; 25: 413–36.
32 Haseman JK. An alternative perspective: a critical evaluation of the Waddell threshold extrapolation model in chemical carcinogenesis. Toxicol Pathol. 2003; 31: 468–70.
33 Crump KS, Clewell HJ. Evidence of a "clear and consistent threshold" for bladder and liver cancer in the large ED01 carcinogenicity study. Toxicol Sci. 2003; 74: 485–6.
34 Haseman JK. Response to Waddell & Rozman. Toxicol Pathol. 2003; 31: 715–6.
35 Gaylor DW. Letter to the editor. Toxicol Pathol. 2003; 31: 572.
36 Anderson ME, Conolly RB, Gaylor DW. Letter to the editor. Toxicol Sci. 2003; 74: 486.
37 Slob W. What is a practical threshold? Toxicol Pathol. 2007; 35: 848–9.
38 Wanibuchi H, Wei HM, Karim MR, Morimura K, Doi K, Kinoshita A, Fukushima S. Existence of no hepatocarcinogenic effect levels of 2-amino-3,8-dimethylimidazo[4,5-f]quinoxaline with or without coadministration with ethanol. Toxicol Pathol. 2006; 34: 232–6.
39 Royston P, Altman DG. Approximating statistical function by using fractional polynomial. The Statistician. 1997; 46: 411–22.
40 Goetghebeur EJT, Pocock SJ. Detection and estimation of J-shaped risk-response relationships. J Rad Stat Soc. 1995; 185A: 107–21.
41 Silverman BW. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. J Rad Stat Soc. 1985; B: 1–52.
42 Gennings C, CarterWH Jr, Carchman RA, Teuschler LK,.Simmons JE, Carney EW. A unifying concept for assessing toxicological interactions: changes in slope. Toxicol Sci. 2005; 88: 287–97.
43 Box GEP. Science and statistics. J Am Stat Assoc. 1976; 71: 791–9.
44 Box GEP, Draper NR. Empirical model-building and response surfaces. New York: Wiley; 1987.
45 Draper NR, Smith H. Applied regression analysis. 3rd ed., New York: Wiley; 1998.
46 Crump K. A new method for determining allowable daily intakes. Fundam Appl Toxicol. 1984; 4: 854–71.
47 Crump K, Hoel D, Langley C, R. Peto R. Fundamental carcinogenic processes and their implications to low-dose risk assessment. Cancer Res. 1976; 36: 2973–9.
48 Crawford M, Wilson R. Low-dose linearity: the rule or the exception. Hum Ecol Risk Assess. 1996; 2: 303–50.
49 Lutz WK. Susceptibility differences in chemical carcinogenesis linearize the dose-response relationship: threshold doses can be defined only for individuals. Mutat Res. 2001; 482: 71–6.
50 Montgomery DC. Design and analysis of experiments. 6th ed. New York: Wiley; 2004.
51 Kavlock RJ, Schmid JE, Setzer RW Jr. A simulation study of the influence of study design on the estimation of benchmark doses for developmental toxicity. Risk Anal. 1996; 16: 399–410.
52 Murai T, Mori S, Kang JS, Morimura K, Wanibuchi H, Totsuka Y, Fukushima S. Evidence of a threshold-effect for 2-Amino-3,8-dimethylimidazo-[4,5-f]quinoxaline liver carcinogenicity in F344/DuCrj rats. Toxicol Pathol. 2008; 36: 472–7.
53 Torous D, Asano N, Tometsko C, Sugunan S, Dertinger S, Morita T, Hayashi M. Performance of flow cytometric analysis for the micronucleus assay-a reconstruction model using serial dilutions of malaria-infected cells with normal mouse peripheral blood. Mutagenesis. 2006; 21: 11-3.
54 Asano N, Torous DK, Tometsko CR, Dertinger SD, Morita T, Hayashi H. Practical threshold for micronucleated reticulocyte induction observed for low doses of mitomycin C, Ara-C and colchicine. Mutagenesis. 2006; 21 15–20.
55 Abramsson-Zetterberg L. The dose-response relationship at very low doses of acrylamide is linear in the flow cytometer-based mouse micronucleus assay. Mutat Res. 2003; 535: 215–22.
56 Filipsson AF, Sand S, Nilsson J, Victorin K. The benchmark dose method–review of available models, and recommendations for application in health risk assessment. Crit Rev Toxico. 2003; 33 505–42.
57 Kroes R, Kleiner J, Renwick A. The threshold of toxicological concern concept in risk assessment. Toxicol Sci. 2005; 86: 226–30.
58 Kroes R, Renwick AG, Cheeseman M, Kleiner J, *et al.* Structure-based thresholds of toxicological concern (TTC): Guidance for application to substances present at low levels in the diet. Fd Chem Toxicol. 2004; 42: 65–83.
59 Cheeseman MA, Machuga EJ Bailey AB. A tiered approach to threshold of regulation. Fd Chem Toxicol. 1999; 37: 387–412.
60 Renwick AG. Toxicological databases and the concept of thresholds of toxicological concern as used by the JECFA for the safety evaluation of flavouring agents. Toxicol Lett. 2004; 149: 223–4.

61 European Medicines Evaluation Agency, Committee for Medicinal Products for Human Use (CHMP). Guideline on the limits of genotoxic impurities. CPMP/SWP/5199/02 (June 28, 2006) London. (http://www.emea.europa.eu/pdfs/human/swp/519902en.pdf).
62 Humfrey CDN. Recent developments in the risk assessment of potentially genotoxic impurities in pharmaceutical drug substances. Toxicol Sci. 2007; 100: 24–8.
63 Watanabe M. Threshold-like dose-response relationships in a modified linear-no-threshold model: application of experimental data and risk evaluation. Genes Environ. 2008; 30: 17–24.

Genes and Environment, Vol. 30, No. 4 pp. 150–159 (2008)

Regular article

Theoretical and Experimental Approaches to Address Possible Thresholds of Response in Carcinogenicity[1]

Kirk T. Kitchin[2]

Environmental Carcinogenesis Division, National Health and Environmental Effects Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, USA

(Received August 28, 2008; Revised October 9, 2008; Accepted October 10, 2008)

The determination and utilization of the actual low dose-response relationship for chemical carcinogens has long interested toxicologists, experimental pathologists, modelers and risk assessors. To date, no unequivocal examples of carcinogenic thresholds in humans are known. However, at least 5 examples of thresholds of preneoplastic foci or tumors have been observed in animals. The two largest dose-response studies utilized 20,880 mice (2-acetylaminofluorene) and 7,200 rainbow trout fry (aflatoxins). In both of these studies linear relationships were observed for DNA adducts and for liver tumors. A threshold relationship was observed for 2-acetylaminofluorene induced mouse urinary bladder cancer. Other comprehensive dose-response studies have examined the chemicals 2-amino-3,8-dimethylimidazo[4,5-*f*]-quinoxaline, 2-amino-1-methyl-6-phenolimidazo[4,5-*b*]pyridine and diethylnitrosamine. Taken collectively, the DNA adduct data for these 6 well studied chemicals are fairly linear. The foci and tumor data show either supralinear, linear or threshold curves, making it difficult to generalize. All the 6 studied chemicals cause multiple biological effects including genotoxicity, cytotoxicity and cell proliferation in complex dose and time dependent patterns that are not fully understood. We do know that there are multiple possible biological defenses (at least 7 pharmacokinetic and 7 pharmacodynamic) against the development of cancer. Currently, we have limited scientific and regulatory understanding of chemicals that act simultaneously or sequentially via both linear and nonlinear carcinogenic pathways (genotoxic and nongenotoxic). If an 100% experimental approach is used to elucidate the dose-response of chemicals of dual carcinogenic dose-response properties (linear and non linear), this would require studying 2 or more such chemicals in a large scale coordinated fashion employing at least 1,000 animals, 5 different treatment groups, 7 different study parameters and 8 different scientific disciplines.

Key words: cancer, threshold, dose-response, genotoxic, mutagenic

Introduction

Looking at the impact of cancer upon human health and society, one quickly sees the enormous incidence of disease, lost human potential and death. Indeed cardiovascular disease and cancer are the two major causes of human disease and death. Using data from the American Chemical Society's Cancer Facts & Figures 2008 report, relative to lifelong human probabilities, the birth to death percent chance of developing an invasive cancer is 44.9% in males and 37.5% in females. The birth to death rate for developing invasive cancer is 12.3 % for breast cancer in females and 16.7% for prostate cancer in males. The rate of invasive prostate cancer development in males aged 70 and higher is 13.4%. There is an approximate 100-fold or greater difference in the age specific rates of several invasive human tumors such as colon & rectum and lung & bronchus (of both sexes) and also urinary bladder and prostate cancer in males. Age related prostate changes and noninvasive prostate hyperplasia, foci and tumors are common in older men (1). Hence, there is a saying among urologists that their male patients will either acquire prostate cancer during their lifetime or die of some other cause prior to developing at least prostate cancer in situ.

Society in general and risk assessment organizations and officials have long struggled with several difficult issues in balancing the conflicting goals of protecting the public health from environmental carcinogens and minimizing the costs of environmental regulations. There are many difficulties that confront a risk assessor. First, there are often orders of magnitude between both (a) the concentrations of chemicals in human exposures and animal experiments and (b) the incidences of cancer in human populations (1–2% for the yearly invasive cancer rates in 60 to 69 year old Americans) and animal bioassay experiments (e.g. often 2 to 100%) (Fig. 1 and

[1]Presented at the *International Symposium on Genotoxic and Carcinogenic Thresholds*, Tokyo, July 22/23, 2008.

[2]Correspondence to: Kirk T. Kitchin, Environmental Carcinogenesis Division, Mail Drop B143-06, National Health and Environmental Effects Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711, USA. Tel: +1-919-541-7502, Fax: +1-919-541-0694, E-mail: kitchin.kirk@epa.gov

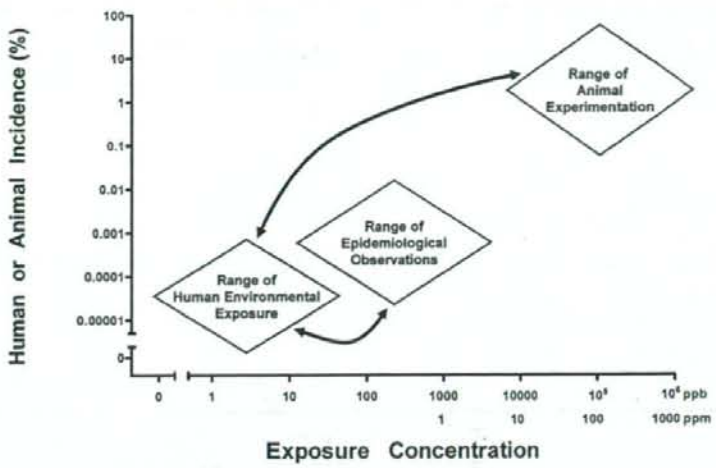

Fig. 1. In human risk assessment we wish to protect a large number of people who are exposed to low concentrations of chemicals. Epidemiological studies and experimental studies in animals can both provide useful and different types of information to risk assessment however the information often comes from quite different ranges of both exposure conditions and incidences.

Table 1. Depending on the chemical of interest and the source of the risk assessment information (epidemiological or animal study), there is often an enormous range of dose or effect extrapolation required to reach the common human exposure concentrations of environmental chemicals (e.g. 2 to $\sim 10^7$)

	SMALL	LARGE
DOSE	2–10	$\sim 10^5$
EFFECT	2	$\sim 10^7$

Table 1). For example, a 50 rodent experiment might at best be capable of telling the difference between a 0 and 2% incidence of cancer, but it cannot provide useful information about human cancer incidences in the range of one cancer case per thousand or one case per million people. At best the limit of sensitivity of common rodent bioassays is about 4–10%. Second, when human epidemiology information is available to assist in assessing risk, the human exposure assessment can never be as complete as in a typical animal bioassay experiment. In addition the human tissue concentration of the parent chemical or important metabolites are rarely known for organs in which cancers frequently develop (e.g. lungs and bronchus, breast, colon and rectum, and prostate). No matter how much information is available and how good it is, a risk assessor is always faced with major uncertainties. Among the largest and most important uncertainties are the true human dose-response relationships in the dose range in which no meaningful animal experiments can be performed.

The term threshold has been used differently by different authors and for different disciplines. In respect to dose-response relationships, there seem to be three major uses of the term threshold. First, the term absolute threshold means that there is no biological effect caused by the exposure at all (2). Second, the term experimental (some people prefer the word practical) threshold means that no biological effect was detected in a particular experiment of a certain sample size and statistical analysis (2). This is somewhat similar to an experiment derived no observable effect level (NOEL). The largest treatment group sample sizes in cancer bioassay experiments is in the range of 2,109 mice (3) and 360 trout (4). Experimental designs for common cancer bioassays use only 50 animals per treatment group. Each individual experimental design will vary in its ability to detect small biological changes and also the degree to which its experimental parameters are correlated with adverse health outcomes such as cancer. The statistical concepts of sensitivity, positive predictivity and concordance are useful in describing the degree to which a particular biomarker is connected to an adverse health outcome (5). Third, the term pragmatic threshold (2) means that a biological effect has occurred but that it is biologically unimportant.

Theoretical Approaches

Moolgavkar originated a theoretical and useful view of carcinogenesis that divides carcinogenesis into the stochastic processes of sequential mutational events and of cell birth, apoptosis, necroses, differentiation and death (6). Figs. 2–4 show schematic versions of a type of multi-stage carcinogenesis model for N required mutations with intervening time for clonal expansion of the number of mutated cells possessing a certain number of mutations. In Fig. 2, the case of purely genotoxic car-

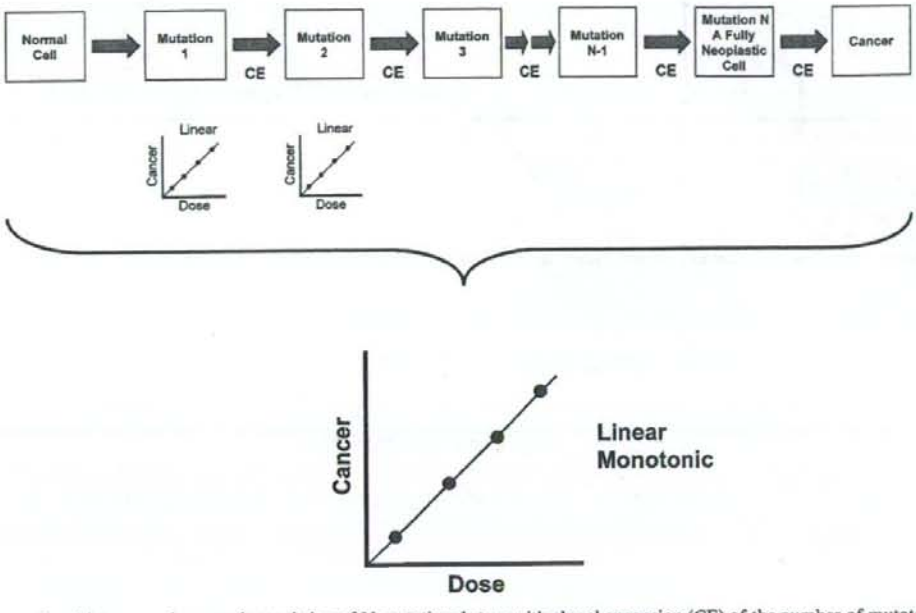

Fig. 2. Scheme of multistage carcinogenesis consisting of N mutational steps with clonal expansion (CE) of the number of mutated cells. In this example the genotoxic carcinogen increases the rates of mutational steps #1 and 2. For strongly mutational compounds the tumor dose-response relationship should be linear at low doses. At higher doses upper limits or asymptotes will be observed near 100% tumor bearing animals.

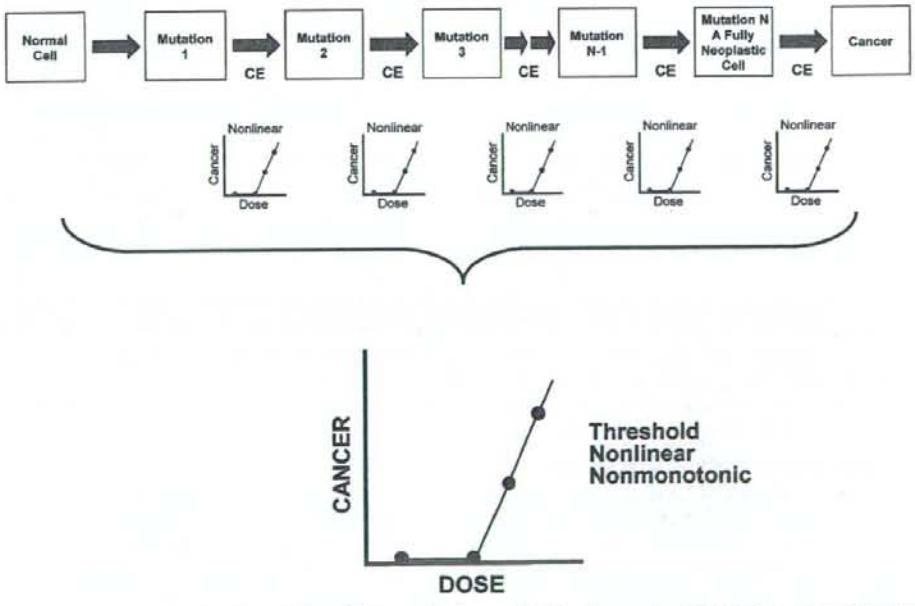

Fig. 3. Scheme of multistage carcinogenesis consisting of N mutational steps with clonal expansion (CE) of the number of mutated cells. In each of the individual clonal expansion steps as well as the overall carcinogenic process there is a nonlinear or threshold type of dose-response relationship.

cinogens which act in a linear dose-response fashion on two mutational steps in the low dose region is presented. Theoretically, if there is no important contribution from cellular kinetics, the best estimate of the dose-

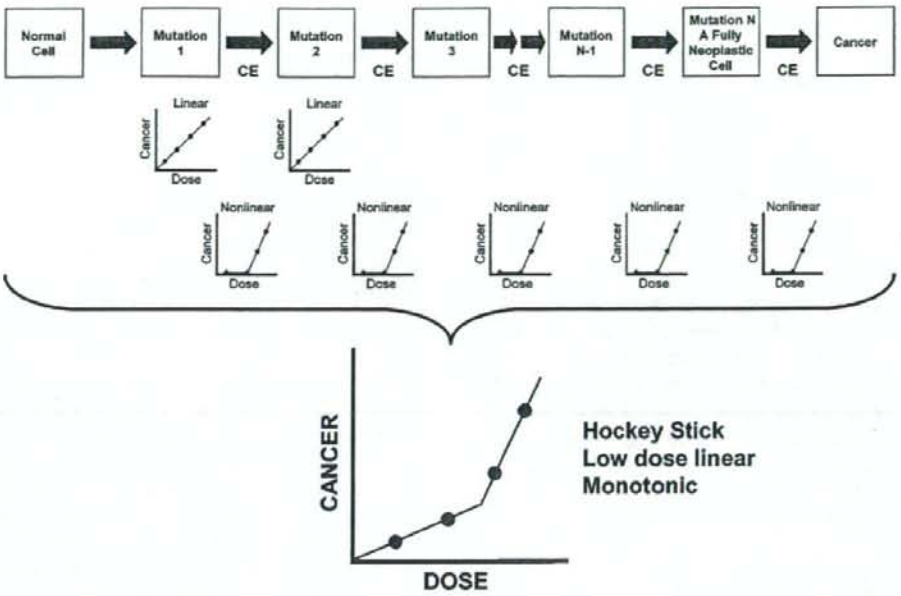

Fig. 4. Scheme of multistage carcinogenesis consisting of two mutational steps increased in rate by dose along with clonal expansion (CE) of the number of mutated cells. This scheme is the mechanistic and mathematical sum of the schemes of Figs. 2 and 3. The dose-response relationship of each of the clonal expansion steps is nonlinear or threshold. The two mutational steps should be linear, the same as in Fig. 2. Overall the relationship between cancer and dose is monotonic, low dose linear and/or hockey stick shaped. At higher doses several processes can contribute to carcinogenicity and upper limits or asymptotes found. At low doses only the mutational process is contributing and the dose-response relationship is linear.

response relationship between cancer incidence and exposure is both monotonic and linear at low concentrations and then at higher doses saturation and asymptote character as the function approaches 100% tumors.

Experiments that sequenced nearly all human genes in 22 cases of human glioblastoma multiforme and 24 cases pancreatic cancer have established that (a) hundreds of individual genes that are mutated in these two cancers and (b) on average 60 genes are altered in glioblastoma (7) and 63 genes in pancreatic cancer (8), respectively. The number of genetic alterations found in the 24 cases of different pancreatic tumors ranged from approximately 30 to 155 (8). These gene alterations included point mutations, amplifications and deletions (7,8), of which point mutations were the most commonly found.

In Fig. 3, a case of a chemical that lacks any linear contribution to carcinogenesis is presented. This is normally considered the nongenotoxic family of chemicals. In this causal scheme all of the driving factors are associated with nonlinear cellular kinetics (e.g. cytotoxicity and regenerative hyperplasia) and there is zero contribution from linear mutational processes. Overall the dose-response of such chemicals is described as threshold, nonlinear or nonmonotonic (Fig. 3). The word monotonic means always increasing or always decreasing.

In Fig. 4 a more interesting case of a chemical that effects both nonlinear and linear causes of carcinogenicity is presented. The contribution from nonlinear processes at high doses makes the slope large in this high dose range. However, at low doses, the dose-response relationship will still be linear even though the slope may be small.

Another theoretical approach to the problem of low dose carcinogenesis is the theory presented by Upton (9) that might be called the incremental exposure-incremental risk or additivity to background theory. This theory states that if chemical-induced biological effects add to already existing background carcinogenic processes, this will necessarily result in a monotonic or linear dose-response relationship (9).

In the history of toxicology, experimental pathology and carcinogenesis, we were first aware of the linear in dose-response, genotoxic and mutational type of carcinogens. Chemicals that are members of this group include ionizing radiation, cigarette smoking exposures, 2-acetylaminofluorene (CAS 53-96-3), aflatoxins, many polycyclic hydrocarbons and nitrosamine carcinogens.

Some of the first instances of nonlinearity in carcinogenesis were found with 2-acetylaminofluorene induced urinary bladder tumors (3), nitrosamine induced esophageal cancers (10), formaldehyde (CAS 50-00-0)