

Biostatistical Issues in the Design and Analysis of Multiple or Repeated Genotoxicity Assays

by Lutz Edler

Tests for genotoxic or mutagenic effects of chemicals have prompted efficient biostatistical methods for the quantification of dose-response data, especially from the Ames Salmonella/microsome assay. A decision about the genotoxicity of a compound is, however, always based on several assays, and results from multiple or repeated genotoxicity assays have to be combined either qualitatively or, even better, quantitatively. The latter problem is considered here, and issues for design and analysis are addressed. General recommendations for designing genotoxicity assays are given. A long-known methodology for combining quantitative parameters from different experiments is updated and other statistical methods suitable for the combined analyses of multiple assays are presented. Some aspects of design and analysis are elucidated on count data from unscheduled DNA synthesis assays.

Introduction

The increasing number of chemicals, their spread into the human environment, and their consumption by humans urges quantitative evaluations of their potential adverse effects. For this reason, short-term tests (STT) have become a widespread biological assay for detecting and assessing genotoxic and mutagenic effects. Growing awareness of genetic factors related to human diseases and the identification of proto-oncogenes and tumor-suppressor genes have sparked renewed interest in the mechanisms of genotoxicity of environmental agents.

Biostatistics has contributed to the design and analysis of genotoxicity assays in important fields: Trend tests have been developed to test for the presence or absence of genotoxic effects, and they superseded multiple pairwise testing. Nonparametric methods replaced parametric ones, suspending the assumption of a Gaussian normal distribution. Transformations were used to deal with variance heterogeneity. Weighted regressions were applied for fitting dose-response models that had been established either as empirical statistical models or as structural mathematical models motivated by biological considerations. Methods for coping with overdispersed data and tests for checking the distributions of the data were developed. Outlier detection and use of historical control information have been established for quality control. Methods for the analysis of a single assay have been summarized recently (1). There have also been suggestions and improvements for the design of genotoxicity assays [See the guidelines of the United Kingdom Environmental

Mutagen Society (UKEMS)] (2). These are mostly intuitive and empirically proved methods rather than theories and they may be called "statistical common sense."

In practice, genetic toxicologists do not conduct only one single assay. Usually, they repeat an assay several times either under identical or varying conditions. This may be done to assure previous results or to cope with the fact that genotoxicity of a compound can be expressed in different ways. The Ames test, for example, has used several tester strains sensitive to different types of mutations. Thus, results from multiple or repeated genotoxicity assays have to be combined somehow. Decision making on the presence or absence of genotoxicity is supported formally by statistical methods of multiple comparisons, and there may be further progress by use of Bayes methods. On the other hand, there is also need for a quantitative combination of results from several assays. Linear models and, more recently, generalized linear models (GLIMs) can be used if the design of the experiment was regular enough. In other cases, long-known methods of weighted means are useful. Their use for genotoxicity assays will be described below. Before dealing with the question of how to combine estimates, some general design considerations for genotoxicity assays are given.

Design and Conduct of Genotoxicity Assays

This section addresses and illustrates basic elements of experimental design. More details on various assays (bacterial and mammalian cell colony and fluctuation, *in vitro* and *in vivo* chromosomal aberration, sister chromatid exchange, Drosophila and dominant lethal) can be found in Kirkland (2). Basic biostatistical elements in designing genotoxicity assays are listed in Table 1. Statistical analysis requires the specification of an end

Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 280, D-6900 Heidelberg, Germany.

This paper was presented at the International Biostatistics Conference on the Study of Toxicology that was held May 13-25, 1991, in Tokyo, Japan.

Table 1. Biostatistical elements in designing genotoxicity assays.

Element	Example
Experimental unit	Cell, cell culture in petri dish, animal <i>in vivo</i> assay
End point	Observability, measurability, identifiability
Conditions for evaluation	Inoculum size, parallel survival assay, incubation prior to treatment, treatment in non-nutrient medium, treatment after growth in nutrient medium.
Treatment groups	Two-sample, many-to-one sample, dose response, controls
Sources of variability	
Sources of bias	
Methods of statistical evaluation	

point, which might be a frequency of counts, a mutation rate, etc. Questions of observability, measurability, and identifiability have to be addressed in some cases, e.g., when a mutation rate has to be calculated from a mutation and a parallel survival assay. The number of cells seeded on a plate (inoculum size) or other conditions of experimentation affect the outcome. The recognition of sources of variability is important. We distinguish within-assay variability and between-assay variability. Within-assay variability contributes to the sampling variation and may be caused by dilution, weighing, pipetting errors, variability in experimental handling, variability of cell division rates in different plates, or counting errors. Between-assay variability contributes to reproducibility and may be caused by physical and chemical properties of the agents, their storage and preparation, or changing growth conditions. More general between-assay variability may be caused by "historical" changes of the protocol, by personnel fluctuation in the staff of the laboratory, or by a genetic drift of the biological material.

A second important aspect is statistical bias: a systematic change of the end point variable, usually to higher or lower values than expected under the ideal experimental conditions. There is no guaranteed protection against biases, but there are possibilities to reduce or at least recognize them by some lessons learned from clinical-trials methodology: Running assays in several laboratories (multicentricity) increases the testing capacity, allows the assessment of interlaboratory variability, and increases representativity of the result. Blind evaluation is possible by using coded chemicals and coded dose groups, and randomization between laboratories and of the order of experimentation might be possible.

Experimentation

Formal experimental requirements have to deal with reproducibility: Physical and chemical properties of test compounds have to be well characterized and controlled. A genetic drift of the biological material during prolonged culturing has to be recognized early. Induction, preparation, and storage of compounds and solvents is a major technical point. Decisions have to be made, for example, between agar-based and liquid-based assays. Use of auxiliary exogenous metabolic activation (e.g., S9 mix in Ames test) or selective agents must be considered because

most target cells have only limited endogenous metabolic capacity. The guarantee of a stable and low spontaneous mutant frequency becomes a major point, when, at the same time, sufficient numbers of cells have to be plated to avoid zero counts. It has been recommended that the number of cells plated initially should assure that a complete set of zero counts occurs with probability not higher than 5% (3). Replicated culturing is basic for statistically evaluable repeated measurements, but a sound statistical estimation of variability requires separate, stable preparation and treatment, not merely splitting the same mixture.

Design

Factors that are a potential source of confounding include number of cells at inoculation, number of replications, number of treatment/dose groups, interval of dosing, and the choice of controls. Mahon et al. (3) recommend a minimum of three dose levels and two separate cultures in each dose group. Blank negative controls and solvent negative controls should be used (4). Positive controls should be incorporated routinely for quality control. Some general requirements have been listed in Table 2. Important elements of statistical design are randomization and assessment of results under blind conditions. The effort and time spent to check for the application of these elements will result in more reliable and reproducible results. Two basic designs can be distinguished: (a) testing for a difference between the treated (exposed) and the untreated controls, (b) establishment of a dose-response relationship with the aim to quantify genotoxic potency. Another benchmark is the choice between parametric and non-parametric models. Although this can still be decided when the experiment is over, it is wise to consider it in the design phase because of its impact on the optimal determination of dose groups and the number of replicates per dose.

End Points and Data Structure

The structure of the data of a genotoxicity experiment depends on the type of the experiment and on the experimental units. Defining factors are shown in Table 3. Mostly, the end point is either a count (e.g., count of revertants, count of aberrations, count of sister chromatid exchanges), or it is a proportion, if the counts have to be related to a baseline number (e.g., the number of surviving mutants among all survivors). Quantitative data are usually hierarchically structured by treatments (dose groups), solvents, replicated cell cultures, and repeated measurements, taken at individual cells. Further stages on top of this hierarchy may be different laboratories or other target cell strains.

Table 2. General requirements of designing a dose-response assay.

Blank and solvent negative controls
Positive control
Minimum of three dose levels
Minimum of two cultures
Increased number of replicates for the negative controls
Number of replicates per dose group depend on the frequency of zero counts: Ames assay, 3-5; sister chromatid exchange, about 50; chromosomal aberration, about 200

Table 3. Defining factors for the data structure of a genotoxicity assay.

Factor	Examples
Agents	Chemicals, radiation, viruses
Experiment	<i>In vitro</i> , <i>in vivo</i>
Measured units	Microbes, cells, cultures, insects/mammals
Design	Treatment/control, dose response
Evaluative criteria	Qualitative, quantitative
End points	Counts, proportions

Sampling Model

The predominant question about sampling models has centered around the appropriate class of statistical distributions for the observed count data. This was triggered by the observation of extra-Poisson variation in the Ames test. Concurrent to methods coping with this so-called overdispersion is the use of transformations toward normality or the analysis of means obtained from an appropriate large number of measurements.

Dose-Response Model

The primary choice of a dose-response model is between a parametric and a nonparametric functional species. Nonparametric methods may be preferred if no agreement on a common sampling model can be found or if one looks for statistical models that are valid under different experimental conditions (laboratory, tester strains, age, and status of test compound). On the other hand, a parametric dose-response model provides an easier way to obtain mutagenic potency measures.

Control of Variability

Weighing, pipetting, transferring microbial cells between vessels and plates, and clumping of cells are factors usually contributing to a high variability. Other sources are varying toxicities on plates, different rates of cell division, dilution or counting errors, variable operators' skills, and calendar time. *In vivo* experiments are further loaded by genetic differences between animals. Use of negative and positive controls is generally advised to control for day-to-day and animal-to-animal variability. Negative control data should lie in an acceptable range and should be compared with historical control data. On the other hand, positive control data should confirm the effectiveness of the entire assay. Table 4 concerns the use of control information in the process of deciding about genotoxicity.

Table 4. Elements of decision making.

Is there homogeneity between negative controls?
How do current negative controls compare with historical negative controls?
How do current positive controls compare with historical positive controls?
Do the treatment groups exceed the negative controls or an absolute threshold?
Do the treatment groups show a dose-response relationship?
Was there increased toxicity or decreased cell survival?
Is there reproducibility between different cultures in the same experiment, between different experiments in one laboratory, or between different laboratories?

Multiple Experiments in Genotoxicology: An Example

Multiplicity of genotoxicity assays is shown clearly in the investigations performed by the U.S. National Toxicology Program (5). Recently, 42 further chemicals were examined using the Ames Salmonella test in four laboratories (up to 3 per chemical), with 3 solvents (up to 2 per chemical), 5 tester strains, 3 S-9 mix categories (none, hamster, rat), and with as many as 4 repetitions per laboratory. This would have led to 1680 assays for one chemical if the maximum number of possible combinations had been used. Of course, most of the possible combinations were not realized because of a reduced number of laboratories, a choice strategy for tester strains, and the choice of the S-9 mixes (and costs). In fact, most chemicals are tested in one laboratory and with one chemical only, which reduces this number to 140 possible combinations. Actually, the total number of dose-response experiments for a genotoxic investigation is usually below 100. For tribromomethane (Bromoform), Zeiger (5) reported 98 dose-response experiments. Reasons for multiple experiments vary. Duplicates are run for confirmation; random inclusion of known positive and negative controls are used for monitoring and controlling the quality of the laboratory; repeated tests are run if unexpected or conflicting results were obtained (6). Table 5 gives the nomenclature for the methods discussed in the following section.

Combination of Estimates

Note that before combining results, it has to be proved that the results are suitable to be combined. This is not easy and may be only partially solvable by statistical tests on heterogeneity or trend. Experimental comparability should be addressed in cooperation with the biologist. On the other hand, there may be situations where one has to come to a conclusion based on a series of estimates if there remain doubts on the comparability.

One Factorial Set of Experiments

Let us consider *I* assays where each has led to an effect estimate, m_i , with a variance estimate, v_i , $i = 1, \dots, I$ [see Cochran (7)]. In some cases we also assume that the estimate v_i

Table 5. Nomenclature for multiple genotoxicity assays.

Group	Dose	Measurements
One assay: dose-response experiment		
Negative control	d_0	$x_{01} \dots x_{0n_0}$
Solvent control		$x'_{01} \dots x'_{0n'_0}$
Dose group	d_1	$x_{11} \dots x_{1n_1}$
.....
.....
Dose group	d_I	$x_{I1} \dots x_{In_I}$
Positive control		$x'_{p1} \dots x'_{pn_p}$
Multiple assays: more than one assay parallel		
Repeated assays: multiple, nonparallel assays		
Experiment: usually an extended study comprising more than one assay		

is based on f_i degrees of freedom and is stochastically independent of m_i . An additive model for the estimate m_i is assumed

$$m_i = m + (m_i - m) + e_i \tag{1}$$

where m is the combined effect, $a_i = m_i - m$ is the interassay deviation, and e_i is the intra-assay error. The variable a_i and e_i are assumed to have expectation 0 and variances σ_a^2 and σ_e^2 . $E[a_i] = 0$ would imply that there is no interassay heterogeneity. Note that this corresponds to a linear model of complete data of the form $Y_{ij} = m + a_i + e_{ij}$, where, for example, a_i and e_{ij} are independent and standard normally distributed with expectation 0 and variances σ_a^2 and σ_e^2 , respectively. Special cases are covariance analyses of either (a) $Y_{ij} = m + \alpha_i + b_j d_{ij} + e_{ij}$ or (b) $Y_{ij} = m + \alpha_i + b d_{ij} + e_{ij}$. Case b contains a combined dose effect, b , which summarizes the single dose effects b_j from case a, if the variances σ_e^2 do not depend on the i th assay. Regression within the groups (assays) leads to an estimate, b , which is a weighted mean of the individual estimates, b_i . Without further assumptions in Equation 1, weighted means of different degrees of complexity can be calculated. A systematic compilation was given in Edler (8) see also Tarone et al. (9). The unweighted mean is the unbiased, least-square estimate of minimal variance as long as inter- and intra-assay variabilities, σ_a^2 and σ_e^2 are equal to 0. Otherwise, a weighted mean estimate with weights $w_i = 1/(\sigma_a^2 + \sigma_e^2)$ is at least of minimal variance. Four classes of means can be distinguished if σ_a^2 and σ_e^2 are unknown, as discussed below.

Unweighted Mean. Variances are given separately for $\sigma_a^2=0$ or $\sigma_a^2 \neq 0$ in Table 6. For the degrees of freedom see Cochran (7).

Grand Mean. Weighting by the degrees of freedom of the variances or by sample sizes, n_i , gives the grand mean (Table 6). Variance and degrees of freedom are obtained similar to the unweighted mean.

Semiweighted Means. Use of weights, $w_i = 1/(\sigma_e^2 + \sigma_a^2)$, is known as semiweighting (Table 6). The variance components, σ_e^2 , will be estimated from the variances, v_i . More difficult is the estimation of the component σ_a^2 . Rao et al. (10) showed four possible solutions, an ANOVA-type variance component estimate, a modification circumventing negative variance components, an unweighted sums of squares, and an MINQ estimate.

Variance-weighted Means. If $\sigma_a^2 = 0$, the weighting reduces to $w_i = 1/v_i$. Then this variance-weighted mean depends heavily on experiments of high accuracy, and assays with a large variance have almost no influence. To counteract this, a so-called partially weighted mean was introduced: The assays are subdivided into a class of low-precision assays weighted by their respective large variances and a class of high-precision assays weighted by a mean of those small variances. Note also the direct correspondence of weighted means and methods of meta-analysis, as well as their relation to Bayesian methods if the choice of a weighting scheme can be related to the choice of a prior distribution.

Multifactorial Set of Assays: Combination of Groups of Estimates

Multiple assays are usually structured by several factors, and it often becomes necessary to combine estimates over some of

Table 6. Unweighted and weighted means.

Type of mean	Equation
Unweighted mean	$\hat{m}_{UW} = \frac{1}{I} \left(\sum_{i=1}^I m_i \right)$
Interassay variability	
Yes ($\sigma_a^2 > 0$)	$\frac{1}{I(I-1)} \sum (m_i - \hat{m}_{UW})^2$
No ($\sigma_a^2 = 0$)	$\frac{1}{I^2} \sum v_i$
Grand mean	$\hat{m}_{Grand} = \frac{1}{\sum n_i} \left(\sum_{i=1}^I n_i m_i \right)$
Interassay variability	
Yes ($\sigma_a^2 > 0$)	$\frac{\sum n_i^2 (\hat{\sigma}_a^2 + v_i)}{(\sum n_i)^2}$
No ($\sigma_a^2 = 0$)	$\frac{\sum n_i^2 v_i}{(\sum n_i)^2}$
Semiweighted	$\hat{m}_{SW} = \frac{1}{W} \sum w_i m_i$
Weights	$Var = \frac{1}{W} \quad W = \sum w_i$
	if $\hat{\sigma}_a^2 > 0 \quad w_i = \frac{1}{\hat{\sigma}_a^2 + v_i} \quad \text{if } \hat{\sigma}_a^2 > 0$
	if $\hat{\sigma}_a^2 = 0 \quad w_i = \frac{1}{v_i} \quad \text{if } \hat{\sigma}_a^2 = 0$
Partially variance-weighted	$\hat{m}_{VW} = \frac{1}{W} \sum w_i m_i$
Increasing order of the variances	$v_1 \leq v_2 \leq \dots \leq v_{I/2} \leq \dots \leq v_I$
Weights	$w_i = \frac{1}{v_i} \quad w_i = \frac{1}{v_i}$
Partial weight for small variances	$v_P = \left(\sum_{i=1}^{I/2} f_i v_i \right) / \left(\sum_{i=1}^{I/2} f_i \right)$

Table 7. Combination of groups of estimates.

Assays	$E_{11}, E_{12}, \dots, E_{1I}$
	$E_{21}, E_{22}, \dots, E_{2I}$
	$E_{R1}, E_{R2}, \dots, E_{RI}$
With estimates	$m_{ri} \quad v_{ri} \quad i = 1, \dots, I, r = 1, \dots, R$

those. It might be reasonable to combine the estimates of the Ames test assays over different metabolic activation levels when there is a small number of repeated assays available for each activation level (Table 7). Combining those groups of estimates can be accomplished stepwise by estimating on each step the basic

parameters and their variances by a semiweighted mean. In a two-step approach, one uses at first the model $m_{ri} = m_r + a_{ri} + e_{ri}$, where m_i denotes the mean effect of the r th group combined over the repeated assays, a_{ri} denotes the interassay deviations in the r th group, and e_{ri} is the error term. In this first step, one can estimate the means for each group as well as their variances. This gives the groupwise pair of estimates \hat{M}_r, V_r . The second step is based on the model $\hat{M}_i = m + A_r + \epsilon_r$, where m is the combined effect and A_r is the intergroup deviation. This second linear model is then the basis for a final semi-weighted mean $\hat{M}_{sw} = (\sum W_r \hat{M}_r) / W$ with $W = \sum W_r$ and $W_r = 1 / (S_{\lambda}^2 + V_r)$, where S_{λ}^2 denotes an estimate of the variance component. If $S_{\lambda}^2 \leq 0$, use of other weights or use the unweighted mean (9) is suggested.

Problems arise if the number of replications is small. Then an ad hoc solution would be a resampling method, where from each group one estimate is sampled randomly and the mean, m_n , of those I values is determined together with a variance estimate, v_n . The random sampling can be repeated many times like a bootstrap procedure. A total mean, m_b , of all repeatedly calculated means, m_n , would give the estimate of the grand effect. A variance estimate can be obtained as the sum of the "bootstrap" variance of the m_n around m_b and a mean variance between the I groups obtained as mean of the variances v_n . For details see Edler (8).

Example: DNA Damage Repair Short-Term Assays

Unscheduled DNA synthesis [UDS (11)] is a type of short-term test that uses the fact that specific cells (e.g., human fibroblasts) are able to synthesize DNA beyond S-phase, between phases G₁ and G₂, (12). UV-induced synthesis of DNA between G₁ and G₂ suggests repair of damaged DNA. In fact, most cells incorporate ³H-TdR into DNA during all stages of the cell cycle after damage. A distinction between S-phase and non-S-phase is achieved by preexposure labeling, resulting in heavily labeled S-phase cells, and postexposure labeling, resulting in lightly labeled non-S-phase cells representing UDS.

The experimental set up for an *in vitro* UDS assay may be as follows (13): Cells are taken from living tissue, incubated, and grown with antibiotics in medium in tissue culture flasks. Growth should be permitted until confluency to avoid replication nuclei, with enormous ³H-TdR uptake. Next the cells are labeled with ³H-TdR to obtain heavily labeled S-phase cells. Then they are exposed to the chemical carcinogens. They are labeled again, and autoradiograms are taken after washing, fixing, and drying them. Use of radioactively labeled thymidine allows the application of autoradiography. The autoradiograms themselves require developing, fixing, washing, drying, and staining the specimen. This enables one to quantify the repair capacity of cells after some exposure to damaging agents as well as the amount of damage that is assumed to correspond to the amount of repair. More experimental details were found by Cleaver (14), who calculated mean number of grain counts of labeled cells adjusted for background by subtracting a mean of grain counts in fields of equal size outside the cell nucleus.

In vivo UDS in rat hepatocytes as complementary short-term assay to the mouse bone marrow cytogenetic was described by

Margolin and Risko (15). They analyzed the end points, sources of variability, and the role of historical controls.

Autoradiography

To understand the variability of the data obtained by autoradiographic methods, a short description of the method is in order. Basically, autoradiography is a photographic method used to determine the distribution of radioactivity in a specimen containing radioactive material. During autoradiography, the radioactive specimen is placed in contact with a photographic emulsion consisting of grains of silver halide, usually bromide. The photographic emulsion is suspended in a gelatin matrix, almost always coated on a glass plate or a film of cellulose acetate or polyester resin. Ionizing radiation liberates electrons, which initiates a reduction of silver ions into metallic silver at the site where radioactivity interacts with the emulsion. Photographic development enhances the effect catalytically, by reduction of additional silver ions in the immediate vicinity of interaction sites. Unaffected silver ions are removed by a fixing solution. The distribution of metallic silver corresponds to the distribution of radioactivity on the specimen. Experimental variations are possible by type and duration of the contact between photographic emulsion and radioactive specimen. Thus one may distinguish between temporary and permanent contact, using the sprinkling, slapping, dipping, floating, or stripping technique for the establishment of the contact (16). The emulsion is fixed and stained after some exposure and development time. Location and intensity of radioactivity of the specimen is indicated by black spots or grains of metallic silver. The end points of the evaluation are the silver grains made visible by this method and their number per cell nucleus. These grains are evaluated microscopically or by image analysis. The quantitative end point is the number and the areas of the grains. The selection procedure for cell identification and counting per nucleus has to be defined; random selection is preferred and "blindness" should be ensured.

The main source of confounding is the background radioactivity and grains generated by other sources than the experimentally controlled radioactivity. This may be the result of prolonged development of the emulsion, exposure to daylight, radiation effects from laboratory environment or cosmic radiation, pressure, chemography, metal ions, static electricity, and differences in their concentration of soluble bromide ions (17). The presence of background grains poses a problem for the analysis of autoradiographic counts. In dose-response experiments, the background can be subsumed under the control group (dose = 0) as long as background intensity does not depend on the dose. Ishikawa and his coworkers (18) used for a graphic display of a plot of the mean number of grain counts versus the logarithm of the dose. This concept was further developed in Thielmann et al. (13). Among several other transformations investigated, the mean versus log-dose gave qualitatively the best results. Plotting the mean number of grain counts versus the logarithm of the dose, a parameter, G_0 , describing the linear increase of the mean number of grains resulting from a dose increase by the factor of $e = 2.72$ was used as the potency. The simple linear regression has the advantage of allowing a straightforward evaluation of repeated experiments. A normal distribution can be assumed because a large number of cells can be evaluated. An investigation

of individual animal net grain counts for the *in vivo* UDS rat hepatocytes assay revealed that mean net grain counts of two or more animals may be considered as normally distributed (15).

Linear Regression Model for Mean Counts

Data for a UDS dose-response assay are the number of grain counts, Y_{ij} per nucleus j ($j = 1, \dots, n_i$), and dose group i ($i = 1, \dots, I$). The increase of the mean number of grain counts per nucleus with dose is usually concave, suggesting a logarithmic transformation of the dose as discussed above. Toxicity or saturation effects, which are not well understood, may cause a downturn of the dose-response curve at high doses. A recursive step-down procedure was used to cope with this. Let the model

$$y_i = \alpha + \beta \ln d_i, i = 1, 1, \dots, I$$

be given for the mean number of grain counts, eventually after subtraction of the mean of the zero dose group. Then the successive regression equations

$$y_i = \alpha + \beta \ln d_i, i = 1, 1, \dots, I-r$$

are evaluated for $r = 0, 1, \dots, I-3$, and doses d_i are discarded as long as a selection criterion holds, such as the minimum estimated standard error (standard deviation of the residuals). If the procedure stops at $R = R$, the resulting model

$$y_i = \alpha + \beta \ln d_i, i = 1, 1, \dots, R$$

is evaluated by simple linear regression.

Another selection procedure could be based on the method of Simpson and Margolin (19). The slope estimate $\hat{\beta}$ is used as measure of repair capability. This simple linear model for the mean number of grain counts per dose has, compared to more complex adaptive procedures, the advantage that it allows a straightforward evaluation of repeated evaluations and repeated experiments per day, several days, or even several laboratories. Because variance homogeneity might not hold in general, weighted regression methods may be indicated. Note that mean counts, Y_i , are no longer independent when the zero dose mean, Y_0 , has been subtracted. However, the differences are independent of Y_0 , and hence the estimation of the slope and the error of variance are unaffected.

Table 8. Results of UDS determination for 11 selected volunteers from the German xeroderma pigmentosum program.*

Strain	G_0	Variance
S1	2.85	0.004
S2	2.36	0.13
S3	2.18	0.62
S4	3.35	0.57
S5	3.46	0.001
S6	2.43	0.31
S7	2.93	0.75
S8	4.09	0.79
S9	3.26	3.27
S10	4.48	1.73
S11	3.09	4.41

UDS, unscheduled DNA synthesis.

* G_0 values of each assay were obtained by linear regression and results of two to three assays were combined by an unweighted mean to a common G_0 value of each cell strain.

Deviations from dose linearity are observed frequently. A simple device is to use a piecewise linear regression, for example, by distinguishing two dose regions. Table 8 shows the unweighted means of an evaluation of 11 selected strains of volunteers from a large-scale evaluation (13). Slope estimates, G_0 , had been obtained from two to three dose-response assays by linear regression as described above. G_0 for strains S1 and S5 had a high precision in contrast to strains S9, S11, and to some extent S10, which had a low precision because of a high interassay variability. The semiweighted mean over these unweighted means resulted in a combined G_0 of 3.0 for all normal strains with variance estimated as 0.07, whereas the partially weighted mean gave a combined $G_0 = 2.9$ with variance 0.09. Without the three low-precision strains, we obtain a semiweighted mean of 2.9 with variance 0.07 and a partial weighted mean of 2.8 with variance 0.05. The variance component was estimated by the MINQ procedure (10).

Dose-response experiments for the *in vivo* UDS assay were analyzed by Margolin and Risko (15) by a simple linear regression, as long as the dose-response curve showed no down-turn at high doses. If there was such a downturn, a simple quadratic regression was applied. In both cases a measure of mutagenicity was calculated from the estimated regression parameters.

Alternative Methods for UDS Count Data

A nonlinear regression model $E[Y_{ij}] = \mu_i(x_i, \beta) = c_i \lambda(x_i, \beta)$ can be applied if μ or λ can be specified as a structural dose-response relationship. The covariate x_i is able to contain arbitrary factor information, i.e., data from rather general designs can be analyzed this way. If it can be shown that the count data Y_{ij} follow a Poisson distribution, Poisson regression methods can be applied (20). If the dependence of the covariate can be expressed via a link function, the solution is also obtained by generalized linear models (GLIMs).

Engel (21) applied quasi-likelihood methods to the analysis of count data from nested designs. The log-quasi-likelihood $l(\mu, x)$ satisfies the equation $\partial l / \partial \mu = (x - \mu) \times V(\mu)$ where $V(\mu)$ is the variance function. Two types of mean variance relationships have been found to be important for count data: $V(\mu) = \sigma^2 \mu$ or $V(\mu) = \sigma^2 \mu^2$. A design where a random factor, B , is nested within a second random factor, A , was considered as well as a design with two fixed factors, A and B , for data Y_{ij} , $i = 1, \dots, I$; $j = 1, \dots, J$, $k = 1, \dots, K$, satisfying a negative binomial distribution with parameter (α_{ij}, P_{ij}) . The variable α denotes the shape parameter of the hidden Γ -distribution and $p = \theta / 1 + \theta$, where θ is scale parameter of the Γ -distribution. Two cases are considered for the second design: *a*) only α_{ij} depends on the two factors (and θ is independent of A and B), *b*) only θ_{ij} depends on two factors. Case *a* corresponds to a constant mean/variance ratio dependent on the mean. Case *b* can be imbedded into a GLIM only if α is known because otherwise the distribution does not belong to an exponential family.

Conclusions

Biostatistics has made important contributions to an unbiased and efficient analysis of the Ames Salmonella assay; and despite the variety of genotoxicity assays, this methodology seems to be applicable or adjustable to a considerable number of genotoxicity

assays. One has to account for considerable variability of the outcome variable of an assay because of factors acting during the experimental progress as well as because of conditions varying between experiments. Variability can be partially controlled by statistical methods. This necessitates designs with negative and positive controls and use of replicates. Otherwise, common biometric principles of experimental design apply to genotoxicity assays. This includes the comprehensive, formal planning of the whole investigation. Blind evaluation, reference evaluation, and principles of randomization should be established and repeated assays should be planned at the beginning of an investigation. Sequential methodology may be helpful and should be explored further. Assays of a planned investigation should be checked for heterogeneity of distribution of the outcome values (e.g., means and variances). Weighted means are shown in this contribution as an elementary method for combining estimates obtained from genotoxicity assays and provide summary measures. They can be applied stepwise in higher-order designs. Statistical regression models may be applied to well designed factorial experiments and to studies with multivariate covariables.

REFERENCES

- Vollmar, J., and Edler, L. Tabular overview of statistical methods proposed for the analysis of Ames *Salmonella* assay data. In: Lecture Notes in Medical Informatics (L. Hothorn, Ed.), Springer-Verlag, Heidelberg, 1991, pp. 42-48.
- Kirkland, D. J., Ed. Statistical Evaluation of Mutagenicity Test Data. Cambridge University Press, Cambridge, 1989.
- Mahon, G. A. T., Green, M. H. L., Middleton, B., Mitchell I. G., Robinson, W. D., and Tweats, D. J. Analysis of data from microbial colony assays. In: Statistical Evaluation of Mutagenicity Test Data (D. J. Kirkland, Ed.) Cambridge University Press, Cambridge, 1989, pp. 26-65.
- Margolin, B. H. Statistical studies in genetic toxicology: a perspective from the U.S. National Toxicology Program. *Environ. Health Perspect.* 63: 187-194 (1985).
- Zeiger, E. Mutagenicity of 42 chemicals in *Salmonella*. *Environ. Mol. Mutagen.* 16 (suppl. 18): 32-54 (1990).
- Piegorsch, W. W., and Zeiger, E. Measuring inter-assay agreement for the Ames *Salmonella* Assay. In: Lecture Notes in Medical Informatics (L. Hothorn, Ed.) Springer-Verlag, Heidelberg, 1991, pp. 35-41.
- Cochran, W. G. The combination of estimates from different experiments. *Biometrics* 10: 101-129 (1954).
- Edler, L. Zusammenfassung von Schaetzungen aus wiederholten Experimenten und Gruppen von Experimenten durch gewichtete Mittelbildung. Technical Report 6185. Abt. Biostatistik, Deutsches Krebsforschungszentrum, Heidelberg, 1985.
- Tarone, R. E., Scudiero, D. A., and Robbins, J. H. Statistical methods for *in vitro* cell survival assays. *Mutat. Res.* 111: 79-96 (1983).
- Rao, P. S. R. S., Kaplan, J., and Cochran, W. G. Estimators for the one-way random effects model with unequal error variances. *J. Am. Stat. Assoc.* 76: 89-97 (1981).
- Djordjevic, B., and Tolmach, L. J. Response of synchronous populations of HeLa cells to ultraviolet irradiation at selected stages of the generation cycle. *Radiat. Res.* 32: 327-346 (1967).
- Rasmussen, R. E., and Painter, R. B. Evidence for repair of ultra-violet damaged deoxyribonucleic acid in cultures mammalian cells. *Nature* 203: 1360-1362 (1964).
- Thielmann, H. W., Popanda, O., and Edler, L. XP patients from Germany: correlation of colony-forming ability, unscheduled DNA synthesis and single-strand breaks after UV damaging Xeroderma pigmentosum fibroblasts. *J. Cancer Res. Clin. Oncol.* 104: 263-286 (1982).
- Cleaver, J. E. Xeroderma pigmentosum: a human disease in which an initial stage of DNA repair is defective. *Proc. Natl. Acad. Sci. U.S.A.* 63: 428-435 (1969).
- Margolin, G. H., and Risko, K. J. The analysis of *in vivo* genotoxicity data: case studies of the rat hepatocyte UDS and mouse bone marrow micronucleus assays. In: Evaluation of Short-term Tests for Carcinogens. Report of the International Program on Chemical Safety's Collaborative Study on *In Vivo* Assays, (J. Ashby, F. J. de Serres, M. D. Shelby, B. H. Margolin, M. Ishidate, Jr. and G. C. Becking, Eds.), 1988, Cambridge University Press, Cambridge, pp. 129-142.
- Rogers, A. W. Techniques of Autoradiography. Elsevier, Amsterdam, 1973.
- Schulze, B. Autoradiography at the cellular level, Vol. 3, Part B. Physical Techniques in Biological Research, 2nd ed. Academic Press, New York, 1969.
- Ishikawa, T., Takayama, S., and Ide, F. Autoradiographic demonstration of DNA repair in rat epithelium treated with chemical carcinogens *in vitro*. *Cancer Res.* 40: 2898-2903 (1980).
- Simpson, D. G., and Margolin, B. H. Recursive nonparametric testing for dose-response relationships subject to downturns at high doses. *Biometrika* 73: 589-596 (1986).
- Frome, E. L. Regression methods for binomial and Poisson distributed data. In: Multiple Regression Analysis: Applications in the Health Sciences (D. Herbert and R. Myers, Eds.), The American Institute of Physics, New York, 1986, pp. 84-123.
- Engel, J. The Analysis of Dependent Data. Proefschrift, Wageningen, Netherlands, 1987.