Drug Information Journal, Vol. 31, pp. 327–334, 1997

0092-8615/97

RECOMMENDATIONS FOR BIOSTATISTICS OF REPEATED TOXICITY STUDIES*

LUDWIG A. HOTHORN, PHD
University of Hannover, Hannover, Germany
K. K. LIN, PHD
Food and Drug Administration, Center for Drug Evaluation and Research, Rockville, Maryland
C. HAMADA, MS
University of Tokyo, Tokyo, Japan
W. REBEL, MD
Boehringer Mannheim GmbH, Mannheim, Germany

The Drug Information Association "3rd Annual Biostatistical Meeting" was held on August 27 and 28, 1996 in Tokyo. The purpose of this meeting was to discuss biostatistical recommendations for repeated toxicity studies. The purpose, objective, design, conduct, analysis, and interpretation are described.

Key Words: Repeated toxicity studies; Biostatistics; Recommendations

PURPOSE

THE PURPOSE OF repeated toxicity studies is to determine the dose levels (or treatments) that cause toxic effects upon repeated administration of substance to mammals and/or dose levels at which no toxic effects occur. The duration varies from one week to two years depending on the objective and the expected period of clinical use (short term: one month or shorter; subchronic: 3 –12 months and chronic: 12 months and longer [The non-neoplastic findings of a carcinogenicity or combined carcinogenicity/chronic toxicity study can be analyzed in an analogous way]). The administration route also depends on the clinical administration. Normally two species, selected from rodents and nonrodents, are used. Details for conduct, design, performance, and analysis are described in several guidelines, for example, those of the United States, Organisation for Economic Co-operation and Development, European Union, Japan, and others.

On the one hand, repeated toxicity studies are "screening" or pilot studies because of their toxicological objective (general safety/risk assessment), their decision making based on numerous endpoints, and repeated measures. On the other hand, repeated toxicity studies can also be designed in a directional manner, for example, for a specific toxicological mechanism based on an a priori assumption (eg, the substance under investiga-

*The views expressed are those of the authors and not necessarily those of their employers.

Presented at the DIA "3rd Annual Biostatistical Meeting," August 27–28, 1996, Toyko, Japan.

Reprint address: Prof. Dr. Ludwig A. Hothorn, University of Hannover, LG Bioinformatics, Herrenhauser Str. 2, D-30419 Hannover, Germany.

tion belongs to a class of chemical substances with a known toxic mechanism) or results from earlier studies.

Usually, repeated toxicity studies are performed as a battery of tests, for example, four-week, three-month and six-month studies in rats and 6/12-month studies in dogs. Sometimes the information of no observed effect level and other findings such as toxic effects, target organs, and reversibility obtained from intermediate repeated toxicity studies (short-term and subchronic studies) are to be used to design a chronic study. Therefore, the objective of an intermediate study is more of a "screening" or pilot rather than to study a specific toxicological mechanism. From the biostatistical point of view, however, each study should be analyzed independently. Decision making based on all studies is very important, although suitable biostatistical methods are still not available.

From a biostatistical point of view, studies in "small" and "large" animals differ according to the sample size and replication of repeated measurement. Unfortunately, from the viewpoint of balancing the false positive to false negative error rate, the analysis of studies with large animals (small sample size: 3, 4, 5) or even with rodents based on a sample size of less than 10 is difficult. Therefore, this recommendation is primarily directed at studies based on studies with sample sizes larger or equal to 10.

THE OBJECTIVE OF STATISTICS IN REPEATED TOXICITY STUDIES

The biostatistical evaluation of repeated toxicity studies should support the decision of whether a finding is positive or negative and evaluate the magnitude of the toxic effects quantitatively, for example, by:

- Summarizing the data,
- Performing statistical tests,
- Investigating the dose-response relationship,
- Identifying sources of variation,
- Analyzing confounding factors,
- Estimation of reference values of historical controls for characterization species-, laboratory-, sex- and endpoint-specific "normal values,"
- Identifying the correlation between several endpoints taking the many-to-one design into account,
- Identifying outliers under the specific conditions of safety studies, and
- Exploring the possible mechanism of toxicity.

It is obvious that biostatistics plays a basic role in adequate toxicity assessment. The statistics in repeated toxicity studies, however, possess a confirmatory (decision making) and an exploratory (effect description) character simultaneously.

THE EXPERIMENTAL DESIGN

The Study Protocol

The study protocol is an important document in the design and evaluation of a toxicity study. Normally it includes descriptions of the following statistical terms:

- Type of experimental design, for example, a negative control and k dose groups, both sexes,
- Experimental unit, for example, a rat,
- Randomization, for example, method of randomization: unrestricted or restricted,
- Endpoints and their scale type, for example, histopathological finding (dichotomous: number of findings/ number of animals at risk),
- Repeated measurement, for example, a timetable with endpoint-specific measures,
- Other approaches for bias reduction, for example, blinding,
- Statistical analysis: model and hypotheses, tests and other statistical methods, sample size, type I and II error rates, adjustment of multiplicity problems, treatment of outliers, and
- Presentation of results.

Sometimes during the course of a study unpredictable factors are observed. There-

fore, the study protocol should be modified exactly according to such unexpected events. The protocol modifications should be submitted along with the original protocol.

Types of Experimental Design

The many-to-one design, that is, including a concurrent negative control group C–, is generally used in repeated studies. The following design is frequently used: {C–, $Dose_1$, . . . , $Dose_k$}. The objective is to analyze the possible dose-response relationship (dose-response effect or not). In the case of a significant dose-response effect, the highest dose revealing biologically unimportant effects should be estimated (no observed effect dose). This design represents the most common design in repeated toxicity studies.

For special experimental objectives, two other designs are also used:

1. {C–, $Treatment_1$, . . . , $Treatment_k$}, where treatments represent different compounds, fixed drug combinations, application forms, and so forth. The objective is to compare the treatments with a negative control, and
2. A more complex design is further inclusion of a standard treatment (a suitable dose of the clinical standard) or a positive control (known toxic substance at a suitable dose). The objective of a positive control is actual proof of the sensitivity of the animal model. The magnitude of positive control and/or standard treatment can help to quantify toxic effects in a relevant scale.

Usually, the number k of dose or treatment groups included should be kept as low as possible (eg, $k \in \{(2, 3, 4\})$ due to multiplicity problems, homogeneity of the animal model, and minimizing the number of animals. Depending on the type of design, specific statistical procedures should be selected.

This many-to-one design belongs to a factorial design with the factors treatment, sex (males, females), time (eg, week 0, 4, 13, 26), and replicated measurements (eg, hematological endpoints measured twice from one blood sample, or histopathology of paired organs). Usually the many-to-one design is analyzed independently for each level of the other factors (see body mass, mortality). Analyzing this factorial design using analysis of variance models (possibly including covariates), however, can be helpful to increase power.

Sample Size Estimation Versus Use of Guideline-related Minimal Numbers

The sample size of repeated toxicity studies should be determined by a priori defined type I and II error, the underlying variability, the kind of testing hypotheses and methods, and so forth. Repeated toxicity studies, however, are screening studies with multiple endpoints with large differences in endpoint type, variance, and relevant difference. Although sample sizes from regulatory guidelines are not based on a power calculation, they can usually be used for screening studies. If a priori, however, a selected toxic mechanism is to be investigated, for example, nephrotoxicological effect, for the related endpoints, a sample size estimation based on a statistical approach should be used. The same procedure should be used in repeated studies, for example, if in a short-term study one or more endpoints were selected as relevant, the sample size for the following study should be designed based on this information. A critical point for sample size choice is between "small" animals (eg, mice and rats) and "large" animals (eg, dogs and monkeys). From the statistical point, the size of the animals does not influence sample size, but from a viewpoint of decision making this is not the case. Therefore, studies in dogs and monkey with higher relevance to humans should be performed only if much information is available from the other toxicological studies. These studies can then be designed for a few clear experimental questions. The type II error = false negative rate in such studies with minimal sample sizes, however, is high. If the sample sizes are too low, for example, ≤ 3, statistical testing does not seem to contribute to reliable decision making.

Due to statistical arguments (some in-

crease in power) the sample size for negative control will be selected higher than the treatment or dose groups (square root rule). Sometimes a double sample size will be used and randomized to one single negative control group. An alternative consists of the randomization to two negative control groups each of sample size n_i, the so-called dual control technique. Comparing both control groups can give information on reproducibility and/or heterogeneity effects and type II error can be reduced. The use of dual controls, however, might inflate the false positive rate in evaluating categorical endpoints if extra-binomial variation exists. Therefore, statistical tests for the existence of extra-binomial within-study variability should be performed before pooling the data of the dual controls.

Dose Selection

Selection of the dose levels in a repeated study is not simple. Statistical methods can help if earlier studies can be analyzed. Frequently, dose selection is a step-wise procedure.

Number of Dose Groups

For several reasons a design using three dose groups seems optimal. Objectives are possible, however, where fewer or more dose groups are more appropriate (eg, dose finding studies for a carcinogenicity assay).

Experimental Unit

The experimental unit in repeated studies is simply the experimental animal, for example, the rat.

Confounding Due to Housing Conditions

Differences in temperature, light, noise, and so forth can occur in the animals' housing. To avoid these possible confounding effects, it is suggested that cages be rotated systematically. Moreover, individual caging should be used if possible. Group-wise housing or other heterogeneities related to sub-groups, for example, palpation or blood taking, should be avoided to ensure the independence assumption.

Recovery Period

One of the important objectives of repeated studies can be the evaluation of reversibility of toxic effects, especially for toxic effects which depend on the pharmacological main effect. In this case, the recovery period should be designed in a suitable manner: a sufficient number of surviving animals and also investigation of negative control animals (some guidelines recommend the analysis of treatment group animals only). When the sample size is too small, for example, ≤ 3, statistical testing does not seem to contribute to reliable decision making.

CONDUCT

Standardization

Before treatment with the substance the responses of the experimental animals in each group should be as equal as possible. Standardization techniques for the animals according to age, initial body weight, health condition, and so forth can be helpful. For example, inclusion and exclusion criteria can be defined, for example, to select animals according to their body weight before randomization between the predefined interval $[\text{weight}_{\text{lower}}; \text{weight}_{\text{upper}}]$.

Randomization

Randomized assignment of the animal to the experimental groups is the basis for qualitative comparison between the groups for treatment differences. Therefore, randomization must be applied wherever feasible.

Blinding

On the one hand, blinding increases the workload for performing a repeated toxicity study. On the other hand, blinding is one

technique to reduce conscious and unconscious bias. Blinding (eg, cage assignment in the animal house or during slide-reading, eg, peer review) should be applied wherever feasible.

Baseline Measurements

Baseline measurements shortly before the first substance treatment (week 0) should be used for noninvasive and nonburdened measures if possible. On the one hand, heterogeneity between experimental groups before treatment can be detected, while on the other hand, treatment differences at later time points can be analyzed using differences from baseline instead of absolute values (covariance reduction). The burden on small animals, for example, mice, however, may be too large.

Endpoints

The guidelines contain detailed lists of endpoints to be measured. From a statistical point of view they should be categorized as: approximate Gaussian distributed (eg, hemoglobin), non-Gaussian distributed (heavily right skewed, eg, ASAT), dichotomous (eg, number of hyperplasia/number of animals at risk), ordered categorical data (eg, graded histopathological findings for each animal), measured once (eg, liver mass), measured repeatedly (eg, hematocrit at Weeks 0, 4, 13, and 26), measured frequently and repeatedly (body mass, daily), reversible and nonreversible (eg, clinical finding of a bloody nose).

Normally, because of the screening character of repeated toxicity studies a priori definition of primary and secondary endpoints is not generally possible. In selected studies, however, such definition is possible. In this case sample size estimation for an univariate design is appropriate.

STATISTICAL ANALYSIS

Descriptive Statistics

Descriptive statistics, as well as single value documentation, are an important part of the study report. Because of different distributed endpoints, both parametric and nonparametric measures can be reported, for example, mean, standard deviation, median, 25% and 75% quartiles, as well as endpoint-specific sample size. Because of small group-specific sample sizes, more detailed measures of distribution, for example, skewness, should not be used routinely. Graphics can be helpful in clarifying the effects and plotting the raw data; group-wise Box-plots can, therefore, be recommended.

Model-based Versus Model-free Procedures

Statistical procedures assume a priori conditions, for example, concerning the underlying distribution or the shape of the dose-response models. Because of the screening character of repeated toxicity studies such a priori assumptions are seldom reliable. Therefore, methods with less restrictive assumptions should be considered. Restricting the alternative hypothesis, however, can reduce type II error markedly. Therefore, restriction of the dose-response alternative to partial or total order seems appropriate in some studies (see below).

Parametric Versus Nonparametric Procedures

The decision between parametric and nonparametric procedures in safety studies is difficult. In textbooks, nonparametric methods are described as robust against (any) real data situation. Keep in mind that one objective of toxicity studies is to find extreme (= pathological) values, nonparametric methods reduce their influence. Moreover, nonparametric methods are nonrobust in the case of variance heterogeneity and the alternative hypothesis is stochastic order instead of location difference. In the case of small sample sizes, for example, studies based on large animals, asymptotic versions of nonparametric methods reveal a loss in power and should be avoided. Here either exact nonparametric tests or parametric tests can be used. There-

fore, nonparametric methods do not seem to be "the method of choice." Both approaches are equally suitable.

Variance Heterogeneity

In dose-response studies an increase in variance with increasing effects is frequently observed (eg, constant coefficient of variation). Both standard parametric and nonparametric methods, however, assume variance homogeneity. Clear heterogeneity, particularly in the case of unbalancedness (large variance and small sample size), may strongly bias the decision (frequently more than deviation from Gaussian distribution). Therefore, tests of variance homogeneity/heterogeneity should be performed. In the case of clear heterogeneity, the use of a modified test procedure, for example, based on α-adjusted Welch-t-tests in a k-sample design, should be applied.

Data Transformation

Suitable transformations and/or weighted procedures can also be used to fit the data closer to the underlying assumptions. Unnecessary data transformations, however, are discouraged. Robust statistical methods should be preferred that do not necessitate transformation, for example, rank transformation or trimming. According to the multiple endpoint problem in repeated studies, a priori definition of the transformation in the protocol does not seem realistic.

Outliers

Treatment of outliers in safety studies is difficult. On the one hand, one objective of safety studies is to identify extreme individual measured values, in the sense of pathological value. On the other hand, extreme values could be outliers without any relation to the substance effect. In contrast to statistical outlier rejection methods in textbooks, such formal approaches should be avoided in safety studies. Declaring an extreme value as an outlier should only be based on technical or biological reasons. The influence of outliers on decision making can be characterized by analysis with and without the critical data. Nevertheless, as a first step of data analysis, statistical methods for outlier detection, for example, group-wise Box-plots, can be helpful to identify such single extreme values in a huge body of data.

Estimation of Confidence Intervals Versus Powerful Testing Procedures

For decision making, confidence intervals and/or testing procedures can be used. Because confidence intervals provide more information, especially with respect to biological relevance, they should be preferred. Moreover, confidence intervals can be seen as an a posteriori approach on proof of safety. In some instances, however, testing procedures could be more powerful than confidence intervals. Since both approaches have advantages and disadvantages, in a real data problem it is difficult to decide which is better. Therefore, the selection of the method primarily depends on the objective of the study.

Univariate Versus Multivariate Procedures

Repeated studies represent a multiple endpoint problem with endpoints of several types. Frequently, independent univariate analysis will be performed. Multivariate approaches based on either formal α-adjustment methods (endpoint-specific testing based on α/number of all endpoints) or multivariate tests with strong power loss with increasing number of endpoints should be avoided. Recent multivariate techniques, without this deficit, for example, summary statistics, can help.

One-sided Versus Two-sided Testing

The decision between one- and two-sided testing in toxicity studies is not simple. Two-sided testing should be used in cases where the a priori direction of effects is not known,

for example, increase or decrease in spleen weight. Power is reduced, however, in two-sided testing in comparison with one-sided testing. For some endpoints, for example, mortality rate, a test for an increase seems to be the only question of interest in such a study, even if a decrease is possible. In dose-response analysis, one-sided testing is more appropriate. For each endpoint, however, the kind of hypothesis must be defined a priori in the protocol.

Restriction of the Alternative Hypothesis

In the many-to-one design including dose groups using order restriction (assuming a monotonic increase of the effect with increasing doses) the type II error can be reduced. Because of the screening character the a priori assumption of total order does not seem appropriate, if downturn phenomena are possible. Therefore, either unrestricted procedures, for example, Dunnett's (1) procedure (however, with lower power) or partial restricted procedures, for example, Shaffer's (2) procedure should be used. Both are robust against downturns at high doses. (Note: These publications are examples only, and not necessarily recommended approaches.)

Dose-Response Analysis

Approaches for dose-response analysis can be categorized as model-based and model-free (assuming a restricted alternative hypothesis only, not a specified function). Because of the screening character, the multiple endpoints, and the small number of dose groups $\in$ (2,3,4), the model-based approach should be avoided. If one wants to use a model-based approach to determine the minimum toxic dose, the four parameter logistic model, the segmented parabolic model, or models in the general form $(\beta_0 + \beta_1 \text{Dose}^{\beta 2})$ $\exp(-\beta_3 \text{Dose})$ are indicated. The objective of the dose-response analysis is to reveal a global effect (dose trend/no dose trend) and a local decision, for example, no observed effect dose.

Asymptotic Versus Permutation Tests

Repeated chronic studies are performed with evaluable animals less than 20 per group, sometimes less than 10. With such small sample sizes, asymptotic procedures (nonparametric, dichotomous, etc.) may have low power. Exact (permutation) modifications may overcome this loss of power. Their possible conservative α-behavior, however, must be taken into account. Unconditional versions can be an alternative in this situation.

Adjustment Due to Multiple Comparisons, Multiple Endpoints, and Repeated Measures

In repeated toxicity studies some sources of multiplicity exist, for example, multiple comparisons, multiple endpoints, and repeated measures. A formal multiplicity adjustment, for example, the Bonferroni method, should be avoided. The primary concern of a toxicity study is the control of consumers' risk (type II error). Using formal multiplicity adjustments to ensure experimentwise type I error, however, the type II error is often increased dramatically. To solve this contradiction either no adjustment (all comparisons based on level α-two-sample tests) or multiple comparison procedures which ensure a small increase in type II error should be used. Normally, adjustment for multiple endpoints and/or repeated measures is not appropriate.

Pooling Over Sex

One of the major limitations of repeated toxicity studies is the small sample size problem. Differences between sexes are to be expected for some endpoints (eg, body weight), but not for others. In the second case, after a preliminary test on sex differences, the data can be pooled if homogeneity exists. For a correct statistical approach a test for equivalence should be performed, which is limited to larger sample sizes. Even using the traditional difference testing approach, the pros and cons of such an analysis speak for a

conditional pooling over sex: two-fold sample size (much more power) and possibly a small bias.

Incorporation of Historical Control Data

Historical control data can be used to check the validity of the concurrent study as well as serve as a basis for comparison between statistical significance and biological relevance. The predictive value of historical controls, however, depends on the compatibility of these controls.

INTERPRETATION

The findings of repeated toxicity studies should be interpreted with caution, because of the "screening" or pilot study character and the relatively high type II error rate (false negative rate) due to small sample size, and so forth. Balancing false-positive versus false-negative rates should be tried. Generally, the type I error should be fixed in the protocol and the type II error should be reported (vice versa in the proof of safety approach).

Statistical significance should not be the only criterion for concluding a positive effect. Biological relevance, for example, based on reference values from historical controls, interaction between endpoints, correlation between studies, should be supported by statistical findings. The classification of a finding into negative, equivocal, and positive should, from a statistical viewpoint, be based on the magnitude of the effect, a dose-response effect, reproducibility of the study (eg, comparison concurrent with historical control), and validity (same findings in both sexes, on several occasions, in several species, in other toxicological studies, etc.)

REPORTING

The printed reports should contain raw data, summary tables, incidence tables, statistical result tables, and graphics. Data conditions, missing values, and other particulars should be noted clearly on the print-outs. All raw data should be stored in a computer file in a suitable format and available for electronic submission in the sense of a computer aided new drug application.

INTEGRITY OF DATA AND SOFTWARE

The validity of decision making also depends on the validity of the data and the software used for management and statistical analysis. Standard operating procedures should be used for data handling, analysis, and reporting. Computer software should be appropriate and validated.

REFERENCES

1. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc.* 1995;50:1096–1121.
2. Shaffer JP. Modified sequentially rejective multiple test procedures. *J Am Stat Assoc.* 1986;81:826–831.