# Biostatistical Design and Analyses of Long-Term Animal Studies Simulating Human Postmenopausal Osteoporosis

**Ludwig A. Hothorn, PhD**
*Unit of Bioinformatics, Hannover University, Hannover, Germany*

**Frieder Bauss, PhD**
*Roche Diagnostics GmbH, Pharma Research Penzberg, Penzberg, Germany, and Institute of Pharmacology and Toxicology, Heidelberg University, Mannheim, Germany*

*Using three well-designed experimental studies as illustration, we demonstrate that the biostatistical design and analysis of long-term animal studies simulating human osteoporosis should be analogous to the design and analysis of randomized clinical trials. This principal is in accordance with the recommendations from the International Conference on Harmonisation guidelines concerning statistical principles in clinical trials (1). An important element of biostatistical study design is sample size. The three studies that are described herein used an a-priori sample size estimation for the one-way layout that included controls and several treatment and dose groups.*

*In these k-sample designs, with at least one control group, both the multiple comparison procedure and trend tests within procedures for identification of the minimal-effective dose are recommended. Although p-values in pharmacology are quite common, confidence intervals should be used according to their interpretation for both statistical significance and clinical relevance. The use of one-sided confidence intervals for both the difference and the ratio to control for proving either superiority or at least noninferiority is demonstrated by real data examples. Relevant and relatively straightforward software is available for biostatistical analysis and can also be used to aid design. In summary, referring to published, well-designed experimental studies can help to assist with ensuring the quality of future investigations.*

**Key Words**
*Ibandronate; Osteoporosis; Many-to-one comparison; Minimum-effective dose; Confidence interval for the ratio; Pharmacological study*

**Correspondence Address**
*Ludwig A. Hothorn, PhD, Bioinformatics, Hannover University, Herrenhaeuserstrasse 2, D-30419 Hannover, Germany (e-mail: hothorn@bioinf.uni-hannover.de).*

*Presented at the Third International DIA Workshop on "Statistical Methodology in Non-Clinical R&D," Barcelona, Spain, September 2002.*

## INTRODUCTION

To date, the biostatistical design and analysis of published pharmacological studies has been highly varied. This has not been assisted by the fact that only a few relevant publications exist that recommend biostatistical procedures (2,3), and those that do exist recommend considerably different methodology compared to clinical trials. Surprisingly, neither international recommendations nor guidelines are available for the statistical design and analysis of pharmacological studies.

Some studies do exist, particularly those relating to long-term animal investigations, for which the biostatistical design and analysis is similar to that of a randomized clinical trial (RCT). However, essential differences remain between clinical and experimental studies. An important example is that quantitative decision-making (with an upper boundary of false-positive and false-negative rates) based on measurements from invasive techniques, for example, bone biopsies, cannot be extrapolated from RCTs. In osteoporosis clinical trials, the primary endpoint for investigating therapeutic efficacy is the reduction in fracture rate, as determined by radiographic analyses. Secondary endpoints include bone mineral density in lumbar spine, hip, and long bones. Additional repeated bone biopsies for the purpose of histomorphometric analyses are only performed in rare cases. Likewise, in pharmacological animal studies, some noninvasive analyses can also be performed during the in-life phase. However, numerous other essential, but invasive measurements, such as those for assessing bone architecture, biomechanical analyses of bone strength, and mineral analyses, are only possible after autopsy of the animals.

Therefore, a global conclusion on the efficacy of a new therapy should be based on both RCTs and long-term animal studies. Due to the importance of these preclinical studies, particularly those in late-stage preclinical development, it should be considered essential that they in-

clude analogous criteria for reproducibility, that is, randomization, a priori sample size estimation (powered study), and protocol-based appropriate statistical evaluation. Therefore, we suggest that recommendations from the International Conference on Harmonisation (ICH) concerning statistical principles in clinical trials (particularly E4 and E9) (1,4) be implemented (for discussion see 5).

Moreover, efficacy studies for comparison with a negative control are quite common in pharmacology. In contrast, RCTs on osteoporosis comparing active drugs with placebo are questioned, and designs for demonstrating noninferiority versus a standard treatment are problematic (6). In this article, we describe the statistical process using three large-scale long-term studies in ovariectomized aged rats, which is an accepted animal model to simulate human osteoporosis.

## THE PRINCIPLES OF PRECLINICAL OSTEOPOROSIS STUDIES

Osteoporosis is defined as a disease that is characterized by low bone mass and microarchitectural deterioration of bone tissue, with a consequent increase in bone fragility and susceptibility to fractures (7). The major cause of this disease is cessation of ovarian function after menopause. Estrogen deficiency in mammals increases bone turnover and results in a reduction in bone mass due to an imbalance between bone resorption and bone formation. Ovariectomy (OVX)-induced bone loss in rats, which shares many characteristics of postmenopausal osteoporosis in humans, is the most accepted animal model to simulate human osteoporosis (8). Consequently, studies in OVX rats are requested by health authorities as a prerequisite for the submission of new drugs intended for the prevention and treatment of osteoporosis (9,10,11).

For the present analyses, three large-scale long-term studies that utilized 735 aged rats were performed using ibandronate (Roche Diagnostics GmbH, D-68305 Mannheim, Germany), which is a highly potent, bone resorption antagonizing, nitrogen-containing bisphosphonate. Ibandronate has demonstrated efficacy in a variety of animal models characterized by increased bone resorption (12). Furthermore, in the clinical setting, ibandronate has proven lasting anti-fracture efficacy in regimens with a dosing interval of greater than two months (13). In our three studies, bilateral OVXes were performed in eight-month old female Wistar rats under general anesthesia. Control animals were sham-operated (Sham) without removing their ovaries. The first study was a dose-finding study (19 groups, $n = 15$ initially), in which daily treatment with ibandronate started immediately after OVX (prevention study) for a duration of 20 weeks. In a second 20-week prevention study (19 groups, $n = 15$ initially), various treatment schedules, all resulting in the same cumulative total dose at the end of the administration period, were compared with daily administration. In the third study (11 groups, $n = 15$ initially), continuous or intermittent treatment over 12 months was initiated 10 weeks after OVX, when considerable bone loss was already demonstrated (treatment study). In all experiments, ibandronate was administered subcutaneously with an administration volume of 2 ml/kg. During the therapy-free intervals, the animals were administered subcutaneously with isotonic saline. The controls received subcutaneous isotonic saline daily throughout the course of the three studies. In all investigations, rats were assigned to the different study groups by stratified randomization according to their body weight.

At the end of the experiments, the right femurs were removed and, after appropriate processing, analyzed for femoral X-ray density (pixels) in the distal region as the primary femoral endpoint. Additionally, the right tibiae were removed and prepared for histomorphometrical analyses of trabecular bone mass (bone volume/tissue volume as percentage: % BV/TV) in the proximal metaphyses as the primary endpoint for a biopsy equivalent. Details of the analytical procedures are described elsewhere (14). In all experiments, a baseline control was sacri-

ficed at study initiation in order to obtain information on the parameters prior to any manipulation of the animals. Success of OVX was confirmed at necropsy by failure to detect ovarian tissue and by the weight of the resected uteri. Animals with a uterus weight of above 350 mg were regarded as not completely ovariectomized and were excluded from the analyses. As a result of this, and on the basis of technical grounds, group sizes for final analyses were n = 11–15 in the two prevention studies (15) and n = 10–15 in the treatment study (16).

## STATISTICAL DESIGN

The use of a control group is highly recommended for demonstrating efficacy in pharmacological studies and it represents the gold standard design. Sometimes, an empty control is used together with a vehicle control to examine the possible influence of the vehicle. If such an effect is unexpected, the use of an empty control should be avoided to limit the number of groups. As a prerequisite in preclinical osteoporosis studies, a sham-operated control should be included to demonstrate the effect of ovariectomy and to exclude any impact of surgery. Additionally, in a pharmacological study, the use of several doses is recommended (at least 2 and at most 6) to demonstrate a dose-response relationship and to select an optimal dose using the concept of minimal-effective dose (eg, 17).

Throughout the course of an experimental study simulating human osteoporosis, numerous measurements are taken at the bone. At baseline, the bone status is characterized by a measurement before ovariectomy and before the drug treatment period. Then, during the treatment period repeated analysis of bone status is performed for the assessment of noninvasive endpoints. Finally, at the end of the study, invasive endpoints are calculated after sacrifice of the animals. As repeated measurements are taken for many treatment groups, the question arises as to which comparisons are relevant. Notably, the main comparison for assessing therapy effect should be between the dose groups and the OVX-control at equivalent time points. In addition, the model of osteoporosis should be proven by comparing baseline versus OVX-control before treatment, OVX versus sham control, and sham versus OVX-dose groups at the same time points, for example, by one sided t-tests.

Here we used one-sided hypotheses throughout. In late phase development, the direction of effect is a priori known in a pharmacological study. However, the number of available animals which can be reliably handled in a single experimental setting is rather limited and should also be restricted according to animal protection purposes. One-sided testing guarantees an appropriate level of power. The efficacy endpoints in clinical trials are tested two-sided or one-sided at $\alpha/2$. One argument for that is the implicit guarantee of an appropriate level of power for the safety endpoints, for which clinical trials are commonly not a priori powered.

Of current interest in osteoporosis is the effect of intermittent versus continuous administration of bisphosphonates. Less frequent, intermittent dosing may provide a more convenient option compared with daily dosing, potentially enhancing compliance. Studies are currently examining whether intermittent administration is equivalent to continuous dosing. These investigations aim to prove at least noninferiority of the intermittent regimen compared to the continuous regimen.

Another aspect to consider when designing a pharmacological study is whether to include a currently licensed, international standard drug (in an optimal dose) as a control. This would unequivocally enable demonstration of proof of effectiveness of the test compound and, in addition, would optimize the selection of the dose for use in future development, which should be at least equivalent to the standard or age-matched sham-controls. Appropriate statistical approaches for analyzing such a design are described by Bauer et al. (18). However, this special design was not applied to our three studies for several reasons. First, a laboratory has the capacity to treat only a certain total number of an-

imals homogeneously. Furthermore, the experimental aim was not to test whether the new drug would be comparable to a standard drug, but to look for an appropriate dose that produced a comparable effect to the respective age-matched healthy sham-controls. Finally, the selected dose range for the dose-finding study was roughly calculated from another study (19), where the relative in-vitro and in-vivo potency of ibandronate compared with other bisphosphonates was elucidated.

It should be noted that a one-way design with females only is used in clinical studies examining the effect of a pharmacological agent in postmenopausal osteoporosis. However, normally in preclinical studies, both genders are considered independent.

Sample size is another important factor in the design of an experimental study from a biostatistical perspective. Commonly, sample size in pharmacological studies is determined by practical limitations. However, such unstructured studies do not allow quantitative statistical reasoning and leave the false-negative rate uncontrolled. To demonstrate this quantitatively, the impact of different sample sizes on the false-negative rate is shown in Table 1 for a comparison of three doses versus a control, a normal distributed endpoint with a ratio of detectable difference $\delta$ to root of variance $\sigma$ of 1 or 2 in at least one dose, one-sided multiplicity-adjusted tests, and a false-positive rate of 0.05.

TABLE 1

**Sample Sizes Needed to Guarantee Some Level of False-Negative Decision Rate**

| | False-Negative Rate | |
|---|---|---|
| $n_i$ | $\delta/\sigma = 1$ | $\delta/\sigma = 2$ |
| 3 | .887 | .635 |
| 5 | .777 | .283 |
| 7 | .667 | .106 |
| 9 | .562 | .036 |
| 11 | .467 | .001 |
| 13 | .380 | < .001 |
| 15 | .309 | < .001 |

According to the ICH guidelines, a maximum false-negative rate of 0.20 should be accepted. Therefore, for studies with either a high variance or a small detectable difference between investigational compounds, sample sizes of 15 or larger per group are necessary. This is only practically possible for certain well-funded pharmacological studies. However, more realistic sample sizes, for example, $n_i = 6$, are only appropriate if a detectable difference of two-fold the standard deviation or even more is accepted. To the extreme, in studies with very small sample sizes, for example, of three or less, it is very difficult to show significance.

Thus, in larger sample size studies, for example, $n_i = 30$, it is simpler to show significance. Often in these larger studies, statistical significance is even observed for biologically nonrelevant parameters. The debate on the difference between statistical significance and biological relevance is endless (20). In the ICH guidelines for RCTs, this problem was solved using Neyman-Pearson testing theory: define a primary endpoint (taken its scale into account), estimate its variance (eg, from historical studies), define an upper bound of the false-positive decision rate (commonly $\alpha = 0.05$) and an upper bound of the false-negative rate (in preclinical studies $\beta \leq 0.30$ seems sufficient), define the direction of decision (one- or two-sided), define the kind and number of comparisons, and then calculate the necessary sample size $n_i$.

In our three osteoporosis studies, many different comparisons for difference or equivalence were performed. To accurately assess these comparisons, we determined the sample size a priori based on the primary endpoint (BV/TV). A one-sided comparison of the six dose groups versus control, without the assumption of order restriction (increasing effects with increasing doses), was selected to demonstrate a minimum increase in the primary endpoint of 10% ($\delta$ for preventive therapy), in at least one of the doses. Based on previous studies we assumed that $\sigma$ = 8%. According to the sample size expression for many-to-one comparisons (21), the sample size

$| n_i | = \frac{2\sigma^2}{\delta^2}(t_{k,R,df=\infty,1-\alpha} + z_{1-\beta})^2$, where $t_{k,R,df=\infty,1-\alpha}$ denotes the $100(1-\alpha)$-th percentile of the multivariate normal distribution with the correlation matrix R for the many-to-one contrasts and $z_{1-\beta}$ denotes the $100(1-\beta)$-th percentile of the normal distribution. Assuming $\alpha = 0.05$ and $\beta = 0.30$, sample sizes of $n_i = 11$ are recommended in this randomized one-way layout. This can be simply calculated by a SAS program for the balanced design:

```
data sample;
alpha=0.05;beta=0.30;sigma=8;delta=10;
  doses=6;
n=2*(sigma**2/delta**2)*(probmc('Dunnett1',.,
  1-alpha,.,doses)+probit(1-beta))**2;
```

Sample size estimation for identification of the minimum-effective dose is also available (22) as well as for testing equivalence (23). However, sometimes, information is not available from previous studies on the variability, primary endpoint, or appropriate dose levels. Adaptive designs can be used, for example, where an internal pilot study is used for obtaining these data in order to design the main study. However, this technique is limited to short-term studies, as shown for an acute pharmacological study by Hothorn and Martin (24).

One considerable difficulty to overcome is when it is necessary to analyze many multiple variables that include some noninvasive repeated measures and some invasive endpoints. To compound the problem, often the parameters differ in distribution (normal distributed, highly skewed, count data, etc.) and variability. Even in a clinical trial, the complex analysis of multiple endpoints, which takes their multiplicity and correlation into account, can rarely be found. In clinical trials, one primary endpoint will often be defined in advance. This is usually inappropriate in an animal pharmacological study where independent univariate analysis of many endpoints can be the only approach. This entails accepting an increase in the false-positive rate. This unwanted effect should be limited by careful interpretation of the results; the final conclusion should clearly discuss significant endpoints that are representative of the same biological effect as one latent variable. On the other hand, if clinically-relevant endpoints show p-values between 0.05 and approximately 0.15, but their variability is clearly larger than that of the 'design-endpoint' and the other significant endpoints, then they should be noted.

## STATISTICAL ANALYSIS

Neyman-Pearson testing theory is commonly used in pharmacological studies and RCTs. Four outcomes of statistical tests are available: the significant/nonsignificant dichotomous decision, categorization into *, **, *** (eg, * if $p < 0.05$, ** if $p < 0.01$, *** if $p < 0.005$), the p-value, or the confidence interval. Although the categorized approach and p-values are frequently cited in pharmacological studies, the use of confidence intervals is considered the most appropriate procedure to use. The two confidence interval values, that is, upper and lower limits (or one limit for one-sided hypothesis), contain information on the significant/nonsignificant decision (ie, exclusion/inclusion the value 0), the directional decision in the case of significance (ie, increasing effect if $CI_{lower} > 0$, decreasing if $CI_{upper} > 0$), the variability ($CI_{upper} - CI_{lower}$), and the magnitude of the significance (distance $CI_{lower} - 0$ for increasing effects). The confidence limits can be directly interpreted in the measurement scale of the pharmacological endpoint, for example, in % BV/TV. This allows a direct interpretation in terms of (pre-)clinical relevance (20). Moreover, the confidence interval can be used for an a-posteriori proof of equivalence (see the example below). In contrast, the p-value is a probability between 0 and 1, with a heavy-tailed distribution only.

The following randomized parallel one-way design [Control, Sham-control, $D_{low}$, $D_{med}$, $D_{high}$, $D_{med\_intermitt}$, $D_{high\text{-}intermitt}$] is used for statistical analysis. In such a k-sample design many-fold use of t- or unadjusted Wilcoxon rank- sum tests is inappropriate, because it violates an experiment-wise type I error. However, the very strict all-pairs comparisons procedure according to Tukey (25) controls the experiment-wise type I error, but with an unacceptable

increase of false-negative rate. Moreover, tests for both efficacy, for example, comparing OVX dose groups with the control and for equivalence, for example, comparing continuous versus intermittent bisphosphonate administration, are performed within this design. Therefore, a priori, that is, in the study protocol, several objectives and their testing procedure were defined.

**COMPARISON 1**

This comparison tested the effect of ovariectomy by comparing the OVX control with the sham control before drug administration by a single two-sample analysis. Further testing was only performed if a clear effect of ovariectomy was demonstrated; otherwise the study is inappropriate for this endpoint. If the p-value of the one-sided t-test for both considered endpoints, femoral X-ray density and the trabecular bone volume of the proximal metaphyses in the tibiae (BV/TV in %), is smaller than 0.0001, that is, further comparisons are possible.

**COMPARISON 2**

Testing for pharmacological efficacy was achieved by comparing the control with all ibandronate dose groups by the many-to-one procedure (26). This assumed approximate normal distribution and variance homogeneity. The box-plot in Figure 1 supports these approximate assumptions.

The study was designed to assess an increased trabecular bone volume relative to the tissue volume (BV/TV) in the ibandronate groups versus OVX control (see Figure 3 in 15). Therefore, one-sided comparisons for an increase in BV/TV were performed. The multiplicity-adjusted p-values and the lower 95% confidence intervals for the difference compared with control are shown in Table 2.

From both the p-values and the lower confidence limits, it can be seen that all doses were significant, with the exception of the lowest dose (0.0001 mg). However, using the confidence interval approach, a clear plateau was observed for all doses ≥0.001 mg/kg/d. These doses induced a ≥11.5% increase in bone volume per tissue volume with 95% confidence probability. With this estimate, the pharmacologist can decide on the direct relevance. We used the SAS procedure GLM (or MIXED) for this analysis and found the syntax to be quite straightforward:

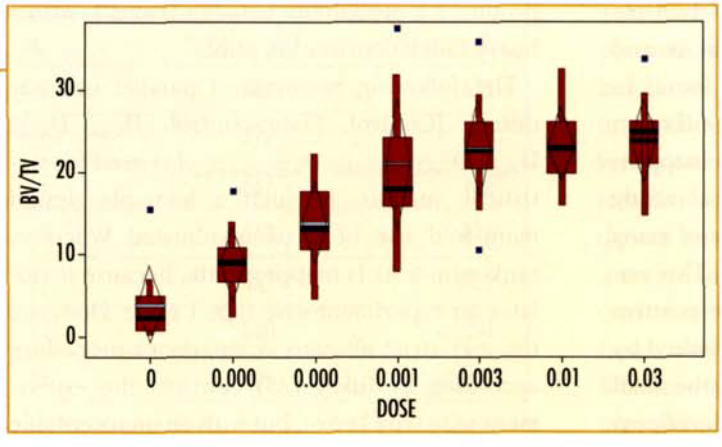
```
proc glm data=xxx;
class dose;
model bvtv=dose;
lsmeans dose/pdiff=controlu cl;
run;
```

**COMPARISON 3**

Testing a significant dose-response relationship seems to be relatively simple. Frequently, either linear regression models (after posthoc dose transformation) or a nonparametric trend test (27) are used. However, both approaches are insensitive to various shapes of the dose-response curve. This is important as the shape is related to the outcome of the study. Two approaches that are sensitive to any shape of the curve are available: the likelihood ratio tests under order

**FIGURE 1**

*Group-wise box-plots for the endpoint BV/TV (dose in mg/kg/d).*



restriction (28) or multiple contrast tests (29). The last approach is based on the maximum test over all contrasts, each sensitive for a possible shape: $T_{max} = \max(T_1, \ldots T_q)$, where each single contrast $T_i$ is sensitive to a particular shape (for an example see 29). The distribution of this $T_{max}$-test is either multivariate t (30) or empirically by a resampling approach, for example, using SAS PROC MULTTEST. A simpler procedure is the use of step contrasts according to Hirotsu (31), which considers the maximum overall possible steps in the dose-response. The syntax of PROC MULTTEST is quite uncomplicated, and the contrast definition for one control and six dose groups makes the step contrast approach clear.

```
proc multtest boot n=10000;
class dose;
test mean(xray/upper);
contrast '1' -6 1 1 1 1 1 1;
contrast '2' -5 -5 2 2 2 2 2;
contrast '3' -4 -4 -4 3 3 3 3;
contrast '4' -3 -3 -3 -3 4 4 4;
contrast '5' -2 -2 -2 -2 -2 5 5;
contrast '6' -1 -1 -1 -1 -1 -1 6;
run;
```

PROC MULTTEST calculates the p-value for each contrast, adjusted for the maximum, and the decision based on the minimum p-value (= maximum test). In this data example, the trend is strong and all p-values are <0.001. However, the p-value for a global trend test is not the only pharmacological question. The minimum-effective dose is also of interest. Under order restriction for designs with a control a step-down procedure can be simply used, starting with the global trend test and eliminating the highest dose as long as this trend was significant (17). In these data, step-down testing up to the lowest dose is possible, that is, the minimum-effective dose for the endpoint X-ray density is 0.0001 mg/kg/d.

**p-values and Confidence Limits for Comparisons Versus Control for the Endpoint BV/TV**

TABLE 2

| Comparison versus Control | p-value | Lower CI in % |
|---|---|---|
| 0.0001 mg | 0.106 | –0.86 |
| 0.0003 mg | 0.0002 | 4.4 |
| 0.001 mg | <0.0001 | 11.5 |
| 0.003 mg | <0.0001 | 13.4 |
| 0.01 mg | <0.0001 | 13.6 |
| 0.03 mg | <0.0001 | 14.4 |

For the X-ray pixel density (see Figure 1 in 15) the Box plot indicates this dose-response relationship in Figure 2.

**COMPARISON 4**

In this comparison, therapeutic equivalence between groups of the same cumulative total dose but with different administration time schedules was analyzed. An interesting pharmacological question is whether the same total dose of a bisphosphonate, administered either daily or

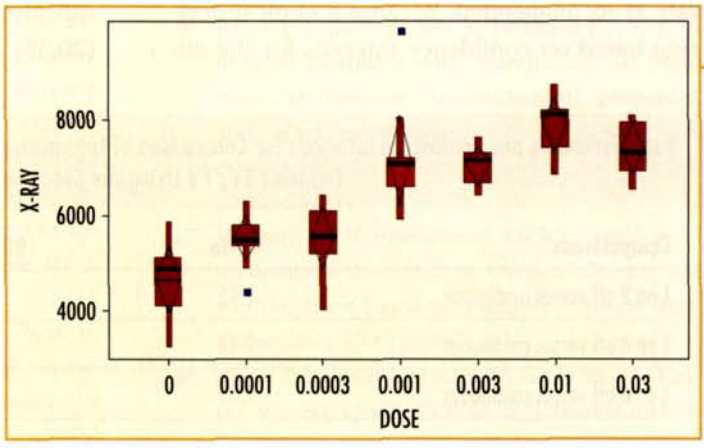
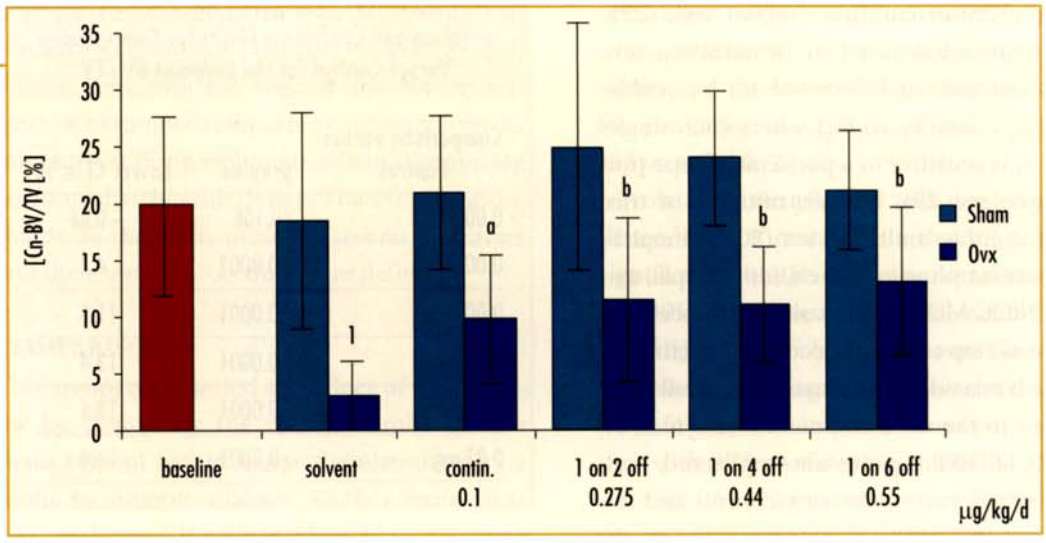



FIGURE 2

*Group-wise box-plots for the endpoint X-ray density (dose in mg/kg/d).*

FIGURE 3




*Effect of the suboptimal dose of ibandronate on trabecular bone mass (BV/TV,%) in tibiae (proximal metaphysis) of ovariectomized (OVX) or sham-operated (Sham) rats (mean ± SD, N = 10–15/group). Significance between sham versus OVX solvent controls: $^{1}p \le 0.0001$; dose schedules versus OVX controls: $^{A}p \le 0.05$, $^{B}p \le 0.005$ (from Bauss [15]).*

intermittently, results in similar effects. Related to the clinical use of bisphosphonates, the statistical question we tested was whether an intermittent regimen is at least noninferior in comparison to continuous administration, or even superior. In one of our studies, this important question was investigated for the optimal and suboptimal total dose and by comparing continuous versus three intermittent ibandronate schedules 1 on 2, 4, and 6 off. Here we present the analysis performed for the endpoint BV/TV, for the suboptimal dose.

A multiplicity adjustment was not used for testing noninferiority. Defining thresholds of acceptance (equivalence margins) can be a serious problem when using equivalence or noninferiority for therapeutic endpoints (32,33). In clinical trials this is difficult; in pharmacology it is almost impractical. Therefore, a posteriori definition, by estimation of confidence intervals, is recommended. Because a clinical decision based on confidence intervals for the difference is complicated (34), we used the confidence intervals for a ratio. Thus, percentage changes to the continuous administration are estimated. BV/TV is approximately normally distributed, therefore, intervals according to Fieller (35), can be defined by means of a simple SAS macro (this macro can be obtained from the first author). The ratio estimates in Table 3 indicate a superiority of intermittent compared with continuous administration. However, the lower $(1 - 2\alpha)$ confidence interval according to Schuirmann (36) and Chow and Shao (37) values are less than one. Thus, an equivalence of both administration schedules can be concluded if a 75% effect is still accepted. This is in concordance with Kanis et al. (6); they describe even a 30% margin as acceptable. A related nonparametric test has recently become available (38). The confidence intervals for the ratio to control as a many-to-one comparative method can also be used for analyzing noninferiority (20,39) and are considered to be a valuable al-

TABLE 3

**Ratio Estimates and Confidence Intervals for Comparison of Intermittent Versus Continuous Administration for the Endpoint BV/TV Using the Suboptimal Dose**

| Comparisons | Ratio | 90%lower CI | 90% upper CI |
|---|---|---|---|
| 1 on 2 off versus continuous | 1.15 | 0.76 | 1.79 |
| 1 on 4 off versus continuous | 1.13 | 0.79 | 1.63 |
| 1 on 6 off versus continuous | 1.31 | 0.90 | 1.98 |

ternative to Dunnett's procedure in proof of efficacy in the second comparison problem.

## CONCLUSIONS

The design and analysis of long-term pharmacological studies, particularly those conducted at a late phase of development, should be conducted analogously to clinical trials according to the ICH guideline. Commonly, k-sample designs with at least one control group are used and, therefore, both multiple comparison procedures and trend tests within procedures for identification of the minimal-effective dose are recommended. Although p-values in pharmacology are quite common, confidence intervals for the difference, which are sometimes more appropriate for the ratio, should be used. This is particularly the case if both efficacy and equivalence (noninferiority) are tested. Relevant software is available. The sample size for these designs should be a priori estimated. Related algorithms have been published in the last few years.

## REFERENCES


1. International Conference on Harmonisation. *ICH E9 Statistical Principles for Clinical Trials.* Geneva, Switzerland: International Conference on Harmonisation; 1998.
2. Hashimoto T, Yamada M, Maeda H. Statistical considerations for preparation of a study report on pharmacological studies. *Nippon Yakurigaku Zasshi.* 2000;116(1):23–28.
3. Andersen H, Spliid H, Larsen S. Statistical models for toxicity and safety pharmacology studies. *Drug Inf J.* 2000;34(2):631–643.
4. International Conference on Harmonisation. *ICH E4 Harmonised Tripartite Guideline 'Dose-Response Information to Support Drug Registration.* Geneva, Switzerland: International Conference on Harmonisation; March 1994.
5. Phillips A, Ebbutt A, France L. The International Conference on Harmonization guideline "Statistical Principles for Clinical Trials": Issues in applying the guideline in practice. *Drug Inf J.* 2000;34(2):337–348.
6. Kanis JA, Oden A, Johnell O, Caulin F, Bone H, Alexandre J-M, Abadie E, Lekkerkerker F. Uncertain future of trials in osteoporosis. *Osteoporos Int.* 2002; 13(6):443–449.
7. Consensus Development Conference. Diagnosis, prophylaxis and treatment of osteoporosis. *Am J Med.* 1993;94:646–650.
8. Kalu DN. The ovariectomized rat model of postmenopausal bone loss. *Bone Miner.* 1991;15(3): 175–192.
9. United States Food and Drug Administration. *Guidelines for Preclinical and Clinical Evaluation of Agents Used in the Prevention or Treatment of Postmenopausal Osteoporosis.* Rockville, MD: United States Food and Drug Administration; 1994.
10. Committee for Proprietary Medicinal Products. *Note for Guidance on Involutional Osteoporosis in Women.* London, England: The European Agency for the Evaluation of Medicinal Products, Human Medicines Evaluation Unit; 2001.
11. World Health Organisation. *Guidelines for the Preclinical Evaluation and Clinical Trials in Osteoporosis.* Geneva, Switzerland: World Health Organisation; 1998.
12. Bauss F. Ibandronate in malignant bone diseases and osteoporosis—preclinical results. *Onkologie.* 1997; 20(3):204–208.
13. Delmas PD, Recker RR, Stakkestad JA, Chestnut CH, Hoiseth A, Weichselberger A, Huss H, von Stein T, Schimmer R. Oral ibandronate significantly reduces fracture risk in postmenopausal osteoporosis when administered daily or with a unique drug-free interval: results from a pivotal phase III study. *Osteoporos Int.* 2002;13:S15 (Abstract O37).
14. Bauss F, Minne HW, Sterz H, Weng U, Wesch H, Ziegler R. Comparative bone analysis via inflammation-mediated osteopenia (IMO) in the rat. *Calcif Tissue Int.* 1995; 37(5):539–546.
15. Bauss F, Wagner M, Hothorn LA. Total administred dose of ibandronate determines its effects on bone mass and architecture in ovariectomized aged rats. *J Rheumatol.* 2002;29(5):990–998.
16. Bauss F, Lalla S, Endele R, Hothorn LA. The effects of treatment with ibandronate on bone mass, architecture, biomechanical properties and bone concentration of ibandronate in ovariectomized aged rats. *J Rheumatol.* 2002;29(10): 2200–2208.
17. Hothorn LA, Neuhaeuser M, Koch HF. Analysis of randomized dose-finding studies: closure test modifications based on multiple contrast tests. *Biometrical J.* 1997;39:467–479.
18. Bauer P, Röhmel J, Maurer W, Hothorn LA. Testing strategies in multi-dose experiments includ-

ing active control. *Stat Med.* 1998;17(18): 2133–2146.

19. Mühlbauer RC, Bauss F, Schenk R, Janner M, Bosies E, Strein K, Fleisch H. BM 21.0955, a potent bisphosphonate to inhibit bone resorption. *J Bone Miner Res.* 1991;6(9):1003–1011.
20. Hauschke D, Kieser M, Hothorn LA. Proof of safety in toxicology based on the ratio of two means for normally distributed data. *Biometrical J.* 1999; 41(3):295–304.
21. Horn M, Vollandt R. Sample sizes for comparisons of k treatments with a control based on different definitions of the power. *Biometrical J.* 1998;40:589–612.
22. Bretz F, Hothorn LA, Hsu JC. Identifying effective and/or safe doses by stepwise confidence intervals for ratios. *Stat Med.* 2003;22(6):847–858.
23. Hauschke D. A note on sample size calculation in bioequivalence trials. *J Pharmacokinet Phar.* 2002; 29(1):89–94.
24. Hothorn LA, Martin U. Application of adaptive interim analysis in pharmacology. *Drug Inf J.* 1997; 30(2):615–619.
25. Tukey JW. The problem of multiple comparisons. Unpublished manuscript (1953). Reprinted in Braun HI, ed. *The Collected Works of* John *W. Tukey.* Volume 8. New York, NY: Chapman and Hall; 1994.
26. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc.* 1955;50:1096–1121.
27. Jonckheere AR. A distribution-free k-sample test against ordered alternatives. *Biometrika.* 1954; 41:133–145.
28. Robertson T, Wright FT, Dykstra RL. *Order Restricted Statistical Inference.* New York, NY: Wiley; 1988.
29. Bretz F, Hothorn LA. Testing dose-response relationships with a priori unknown, possibly non-monotonic shapes. *J Biopharm Stat.* 2001; 11(3):193–207.
30. Bretz F. *Powerful Modifications of Williams' Test on Trend.* PhD Thesis. Hannover, Germany: University of Hannover; 1999.
31. Hirotsu C. The cumulative chi-squares method and a studentised maximal contrast method for testing an ordered alternative in a one-way analysis of variance model. *Reports Stat Applica Res.* 1979;26:12–21.
32. Hauschke D. Choice of delta: A special case. *Drug Inf J.* 2001;35(3):875–879.
33. Ng TH. Choice of delta in equivalence testing. *Drug Inf J.* 2001;35(4):1517–1527.
34. Roehmel J. Therapeutic equivalence investigations: Statistical considerations. *Stat Med.* 1998; 17(15-16):1703–1714.
35. Fieller E. Some problems in interval estimation. *J Royal Stat Soc.* 1954; B16:175–185.
36. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokin Biopharm.* 1987;15: 657–680.
37. Chow SC, Shao S. A note on statistical methods for assessing therapeutic equivalence. *Control Clin Trials.* 2002;23:515–520.
38. Hothorn T, Munzel U. *Exact Nonparametric Confidence Intervals for the Ratio of Medians.* Technical report. Erlangen, Germany: University of Erlangen; 2002.
39. Hauschke D, Kieser M. Multiple testing to establish noninferiority of k treatments with a reference based on the ratio of two means. *Drug Inf J.* 2001;35(4):1247–1251.