F 20895

INFORMATIK BIOMETRIE und EPIDEMIOLOGIE

IN MEDIZIN UND BIOLOGIE

Offizielles Organ
der Deutschen Gesellschaft für
Medizinische Informatik,
Biometrie und Epidemiologie (GMDS) e.V.

Urban & Fischer Verlag
Verlag Eugen Ulmer Stuttgart

Band 30
Heft 3/1999
ISSN 0943-5581

INFORMATIK BIOMETRIE und EPIDEMIOLOGIE
IN MEDIZIN UND BIOLOGIE

Urban & Fischer Verlag Jena
Verlag Eugen Ulmer Stuttgart

Schriftleitung:
Prof. Dr. Markus Löffler, Leipzig

Inhaltsverzeichnis 3/1999

Redaktion
Schriftleiter (verantwortlich im Sinne des Presserechts): Prof. Dr. M. Löffler Universität Leipzig, Institut für Medizinische Informatik, Statistik und Epidemiologie, Liebigstraße 27, 04103 Leipzig, Telefon: (0341) 9716100, Telefax: (0341) 9716109, e-mail: Loeffler@imise.uni-leipzig.de. Verantwortlich für die Mitteilungen der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V.: Dipl.-Volksw. Th. Banasiewicz, Herbert-Lewin-Straße 1, 50931 Köln,Tel. (0221) 4004-865.
Verlag: Vertrieb und Werbung: Urban & Fischer Verlag GmbH & Co. KG, Niederlassung Jena, PF 100537, D-07705 Jena; Tel. (03641) 626-3; Fax (03641) 626500; e-mail: journals@urbanfischer.de. Herstellung und Anzeigen: Verlag Eugen Ulmer GmbH & Co., Wollgrasweg 41, 70599 Stuttgart, Tel. (0711) 4507-0; e-mail: info@ulmer.de Postscheckkonto Stuttgart 7463-700, Zürich 80-47072, Wien 1083.662 Deutsche Bank AG, Stuttgart, Kto. 14/76878, Südwestbank AG, Stuttgart, Kto. 741371006, Herstellung: Sigrid Wolf, Tel. (0711) 4507-194. Verantwortlich für die Anzeigen: Dieter Boger, Lilyen Sever, Tel. (0711) 4507-144, z. Z. ist die Anzeigenpreisliste Nr. 10 gültig. Anzeigenschluß: am 20. der Monate Januar, April, Juli, Oktober.
Druck: Druckhaus »Thomas Müntzer« GmbH, Neustädter Straße 1–4, 99947 Bad Langensalza, Telefon (03603) 399-0.
Abonnementsverwaltung: SFG-Servicecenter Fachverlage GmbH, Zeitschriftenvertrieb: Barbara Dressler, Villengang 2 D-07745 Jena; Telefon: (03641) 626444; Fax (03641) 626443.
Bezugshinweise: Das Abonnement gilt bis auf Widerruf oder wird auf Wunsch befristet. Die Lieferung der Zeitschrift läuft weiter, wenn sie nicht bis zum 31. 10. eines Jahres abbestellt wird.
Erscheinungsweise: (1999) 1 Jahrgang mit 4 Heften.
Abo-Preise: (1999) 1 Jahrgang 392,– DM; Einzelheftpreis 110,– DM. Alle Preisangaben verstehen sich zuzüglich Versandspesen.
Vorzugspreis für Mitglieder der Deutschen Region der Internationalen Biometrischen Gesellschaft auf Anfrage beim Verlag.
Folgende Kreditkarten werden zur Zahlung akzeptiert: Visa/Eurocard/Mastercard/American Express (bitte Kartennummer und Gültigkeitsdauer angeben).
Bankverbindung: Deutsche Bank AG Jena, Konto-Nr. 6284707, BLZ 82070000;
Indexed in „Biological Abstracts" und „Current Index to Statistics".

Informatik, Biometrie und Epidemiologie in Medizin und Biologie **30** (3),
85–86, ISSN 0943-5581

Editorial

That which is not controversial
is of no particular interest.

J. W. VON GOETHE, *Letters*

The objective of bioequivalence studies is to show that different formulations of a drug product (in the simplest case a new test and a current reference formulation) are sufficiently similar in terms of the extent and rate of absorption. From the closeness of the concentration-time profiles it is concluded that the efficacy and safety characteristics of the formulations are also not noticeably different. For the patient, bioequivalence of drug formulations is of interest in two basic situations: (i) When a patient starts with a treatment for the first time and can chose between different formulations, any choice should have the same consequences with respect to efficacy and safety (*drug prescribability*); (ii) When a patient, who is already titrated on an efficacious and safe level of one formulation switches to another formulation, this should not cause a noticeable change with respect to efficacy and safety (*drug switchability*). There exist three different concepts of bioequivalence. Average bioequivalence refers to equivalence with regard to the means of the distributions. As a generalization, population bioequivalence requires closeness of the full distributions. On the other hand, individual bioequivalence aims at demonstrating that the two formulations are similar for a sufficiently large portion of individual subjects in the population. At the moment, regulatory guidelines only require the proof of average bioequivalence for the approval of a generic drug. However, it is recognized that drug prescribability requires population bioequivalence of the formulations and that for drug switchability it is necessary to establish individual bioequivalence.

As a consequence of the continuing discussions of whether average bioequivalence is sufficient to declare drug formulations as comparable, the Individual Bioequivalence Working Group was established in the Center for Drug Evaluation and Research of the U.S. Food and Drug Administration (FDA). The recently published FDA draft guidance 'In vivo bioequivalence studies based on population and individual bioequivalence approaches' (1997; http://www.fda.gov/cder/bioequivdata) essentially mirrors the explorations of this task force. The move from the concept of average bioequivalence to population and/or individual bioequivalence proposed by this draft guidance would entail considerable changes with respect to study design and statistical methods for planning and analysis of bioequivalence trials.

Since its release there is an intensive and controversial debate about the strengths and weaknesses of the draft guidance. The November 1998 meeting of the Working Group 'Pharmaceutical Research' of the German Region of the International Biometric Society was held under the topic 'The FDA draft guidance on bioequivalence assessment'. It was the purpose of this meeting to review the present approaches of bioequivalence and to discuss the FDA draft guidance from various aspects. This issue of *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* contains the papers of four presentations given at this conference thus enabling interested parties to share in this exciting discussion about recent developments in bioequivalence testing.

The introductory paper of Dr. Martina Elze gives an overview of the current approach of bioequivalence testing and illustrates the shortcomings of the average bioequivalence concept by impressive practical examples. It is shown how these disadvantages can be resolved by the concepts of population and individual bioequivalence. However, it is also pointed out that there are still a number of open questions inherent in these alternative approaches. This gives the entry for the following papers where some of these aspects are addressed more closely.

It is one of the novel features of bioequivalence testing via the concepts of population and individual bioequivalence that bootstrap techniques should be applied. Although the popularity of bootstrap methods has continuously grown in the last years, they are still awaiting their routine application in clinical trials. As a consequence, there is a lack of familiarity with this methodology for many practitioners. The paper of Prof. Iris Pigeot reviews the general idea of resampling and demonstrates how bootstrapping is applied to assess population and individual bioequivalence. The algorithms for generating bootstrap percentile intervals and bias-corrected bootstrap confidence intervals are given and the presented methods are related with the recommendations of the FDA draft guidance. It is concluded that there are several unresolved problems in connection with the application of bootstrap methods in bioequivalence testing.

It is surprising that a regulatory guideline proposes a specific software for evaluation although alternative packages exist that have also the facility to succeed with this task. At least, one would expect that this favoured program would be especially suitable for this purpose and fully developed. The paper of Dr. Günther Pabst describes practical experiences when evaluating trials to assess individual bioequivalence. One part of his article is concerned with the software recommended by the FDA draft guidance, and his findings show that there is no cause for illusions. Furthermore, special features of the repetitive design are given which would replace the standard two-period crossover when switching from average to individual bioequivalence. These include ethical and statistical issues as well as consequences for the conduct of bioequivalence trials.

The concluding paper of Prof. Joachim Röhmel gives comments to various aspects of the FDA draft guidance such as design issues, bioequivalence criteria, metrics, and statistical analysis. Special attention is called to the statistical model underlying the new bioequivalence approaches and its relation to the situation when applying the average bioequivalence criterion in repeated crossover designs.

Summarizing the arguments given in the four articles it becomes evident that there is considerable room for improvement of this draft guidance. It is to be hoped that the current version is relevantly revised before it comes into operation.

Finally, we wish to thank Prof. Löffler (Editor-in-Chief, *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*), Dr. Acker (Editorial Assistant) and the reviewers for their advice and support.

Meinhard Kieser

Karlsruhe, May 1999

Informatik, Biometrie und Epidemiologie in Medizin und Biologie **30** (3),
87–95, ISSN 0943-5581

Bioequivalence Trials – Status and Perspectives

Bioäquivalenzstudien – Stand und Perspektiven

Martina Elze, Henning H. Blume

Summary

The current practice of assessing bioequivalence accepted and used by international regulatory authorities is based on average bioequivalence. Due to several potential drawbacks of this concept two new approaches – population and individual bioequivalence – were developed. In this paper the relation between the requirements of appropriate bioequivalence testing from a clinical and pharmaceutical point of view and the corresponding statistical background is demonstrated.

Keywords

Individual bioequivalence, population bioequivalence, average bioequivalence

Zusammenfassung

Die Methodik zur Bewertung von Bioäquivalenzstudien, die gegenwärtig von seiten der Behörden international akzeptiert und angewandt wird, basiert auf dem Konzept der „Average Bioequivalence". Aufgrund potentieller Nachteile dieser Methode wurden zwei neue Ansätze – die „Population" und „Individual Bioequivalence" – entwickelt. In der vorliegenden Publikation werden die Erfordernisse einer adäquaten Bioäquivalenzprüfung aus klinischer und pharmazeutischer Sicht und die dementsprechende statistische Methodik dargestellt.

Stichworte

Individuelle Bioäquivalenz, Populationsbioäquivalenz

1 Introduction

"Two pharmaceutical products are considered to be equivalent when their concentration vs. time profiles, from the same molar dose, are so similar that they are unlikely to produce clinically relevant differences in therapeutic and/or adverse effects" (BLUME et al., 1995a). Based on this definition of bioequivalence the members of the Bio-International Conference '94 agreed upon the clinical implications of bioequivalence trials are evident. The similarity of the concentration vs. time profiles after administration of

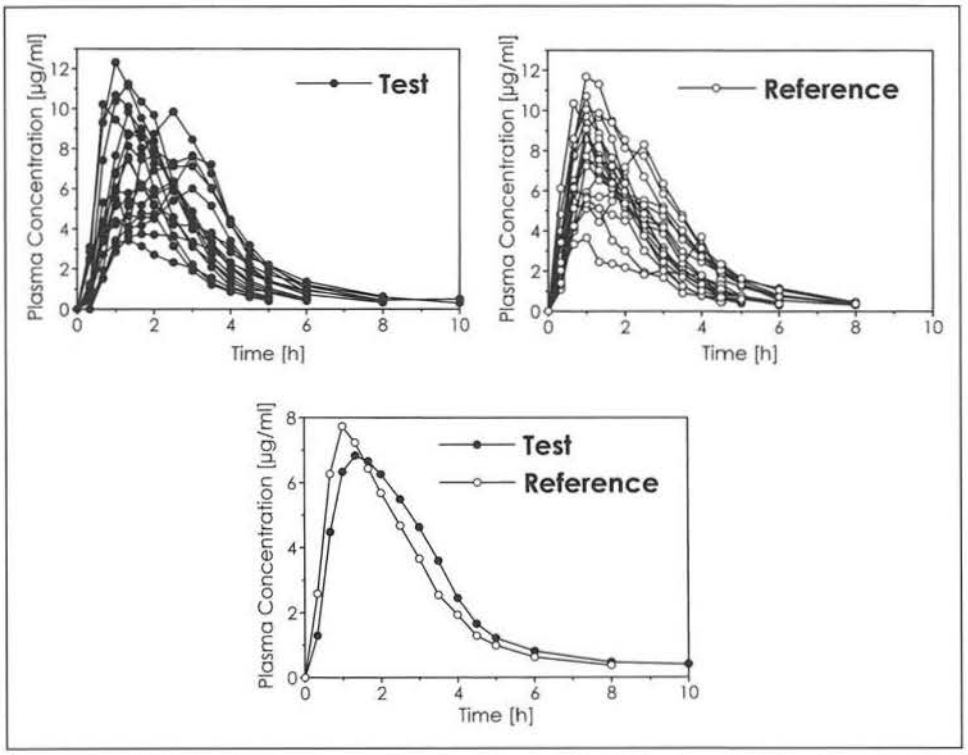


Figure 1: Individual and mean concentration vs. time profiles following administration of the test formulation in comparison to the reference product. 2-period crossover design. 18 subjects. Example for average bioequivalence. Comparable variabilities

two drug products is used as surrogate for similarity in efficacy and safety of the products i.e. for therapeutic equivalence. Bearing in mind these clinical consequences of a positive bioequivalence decision, particular focus on design, study conduct, analysis and especially on bioequivalence criteria and statistical methods for decision is needed.

One of the most important fields for conducting bioequivalence trials is the regulatory approval of generics in comparison to the innovator drug product. Further, bioequivalence studies are performed prior to approval of a new drug product if the investigational product used in clinical trials is different to the final product intended for marketing. The development of formulation and manufacturing parameters as well as clinical investigations are overlapping processes. Changes in formulation (e.g. composition, excipients, dosage form, coat of a tablet) and manufacturing components (e.g. compaction force, particle size of drug substance and excipients, duration, intensity and type of granulation of a tablet) may be necessary in order to obtain the optimal product (MORAIS, 1995; OHM, 1995). Moreover, factors possibly effecting bioavailability (e.g. age, sex, food, drug interactions, hepatic and renal insufficiencies) are usually studied on the basis of a bioequivalence trial. Additionally, assessing bioequivalence is required in case of postapproval major changes in formulation and manufacturing due to e.g. improvements of technology.

The current assessment of bioequivalence accepted and used by international regulatory agencies is based on average bioequivalence. As a consequence of shortcomings of the

conventional concept two new approaches – population and individual bioequivalence – were developed resulting in the draft FDA guidance "In vivo bioequivalence studies based on population and individual bioequivalence approaches" delivered for discussion in October, 1997 (FDA, 1997).
This paper links the requirements of appropriate bioequivalence testing discussed by pharmaceutical and regulatory scientists for years with methodological statistical background provided by biostatisticians. The current practice of average bioequivalence and potential disadvantages will be presented. Further, the concepts of population and individual bioequivalence and how they address drawbacks of the currently used approach will be described.

2 Average bioequivalence

2.1 Current practice of assessing bioequivalence

Currently, two pharmaceutical products – the test formulation compared to the reference product – are administered to a sufficient number of volunteers in a 2-period crossover design. Analytical procedures will provide concentration vs. time data for each volunteer resulting in individual concentration vs. time profiles after administration of the test and reference product (Figure 1). The corresponding mean curves are also depicted in Figure 1. Based on these individual concentration vs. time data, the similarity of the profiles following administration of the test compared to the reference product is to be demonstrated. Next, the concentration vs. time profiles are characterised by appropriate pharmacokinetic parameters, usually by AUC (area under the curve) and $C_{\max}$ (observed maximum concentration) after single-dose administration of immediate release formulations. AUC and $C_{\max}$ are random variables which follow a certain population distribution whose mean is addressed by the definition of average bioequivalence. Test formulation vs. reference product are considered to be bioequivalent if the **means** are "sufficiently similar" with regard to AUC and $C_{\max}$.
According to the present approach sufficiency of similarity is accepted if

$$0.80 < m_T/m_R < 1.25\,,$$

where m_T and m_R denote the population means (AUC, $C_{\max}$) after administration of the test formulation (T) and the reference product (R), respectively.
Following the currently valid FDA and CPMP recommendations of logarithmic data transformation (CPMP, 1991; FDA, 1992) the log scaled criterion is

$$\log(0.80) < \mu_T - \mu_R < \log(1.25)\,,$$

where μ_T and μ_R denote the population means of the logarithmically transformed observations.
The standard method for statistical testing is the two one-sided t-tests procedure (SCHUIRMANN, 1987) usually performed by calculating the 90% confidence interval for the respective ratio ($\alpha = 0.05$). Bioequivalence is accepted if the 90% confidence interval is included in the bioequivalence range 0.80–1.25.

2.2 Potential drawbacks

a. At present, bioequivalence testing addresses only mean ("centre of the distribution") but not variability ("shape of the distribution"). However, AUC and $C_{\max}$ values vary within a subject and between subjects. Variability is due to different reasons e.g.

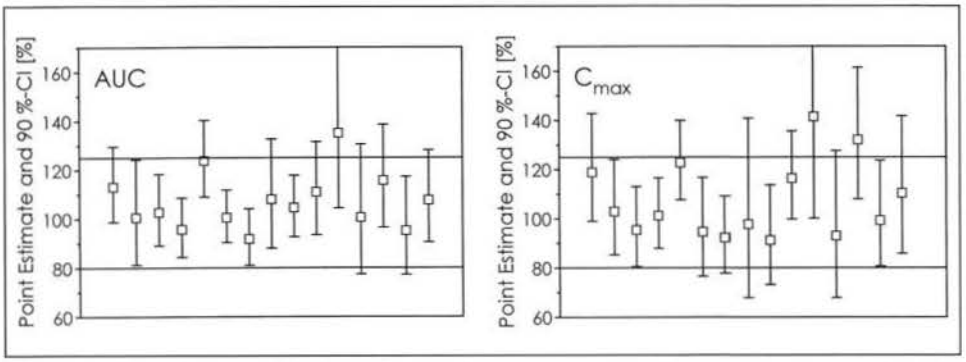


Figure 2: Point estimates and 90% confidence intervals (two one-sided t-tests procedure) for AUC and C_{max} of 15 bioequivalence studies with verapamil immediate release generics of the German market compared to the innovator drug product. 12–18 subjects

differences between subjects, physiological variation within a subject (enzyme activity, gastrointestinal activity), variation in pharmaceutical quality within a batch and variability of analytical methods.

b. The present approach does not address switchability i.e. switching a patient from one to another drug product in case of generic substitution or postapproval major changes.
c. The current practice is based on fixed bioequivalence limits (0.80, 1.25). The practical background will be illustrated by two examples given in the following.

Figure 2 presents point estimates and 90% confidence intervals regarding AUC as well as C_{max} of 15 bioequivalence studies carried out by German pharmaceutical companies with verapamil immediate release generics of the German market compared to the innovator. The studies refer to a 2-period crossover design in which verapamil was administered as single doses (at least 16 subjects). The data were reevaluated according to the two one-sided t-tests procedure (BLUME and MUTSCHLER, 1995b). Based on the traditional limits for acceptance of equivalence (as percentages) the equivalence criterion with regard to AUC was met in only 6 of 15 studies. Considering C_{max}, the criterion was met in only 4 of 15 studies. Consequently, most of the generics could not be considered bioequivalent to the innovator product. Nevertheless, switching a patient from one to another product is frequent therapeutic practice.

This example could provide a supportive argument that the fixed limits may be not appropriate for each drug and formulation.

Validity of the current practice of assessing bioequivalence can be investigated by replicate administration of the reference product. Immediate release products from the German market containing verapamil, thioctic acid, nifedipine and ibuprofen were administered as single doses in at least 16 subjects according to a replicate design i.e. products of the same batch were administered twice. It is expected that the probability of acceptance of bioequivalence of the reference product compared to itself would be close to 100%. Statistical analyses by the two one-sided t-tests procedure using the fixed limits 0.80 and 1.25 resulted in probabilities of a positive equivalence decision with regard to AUC and C_{max} shown in Figure 3. Considering the verapamil reference product as an example, the probability of acceptance of equivalence for AUC is only 40%. The situation concerning C_{max} is worse with a power of merely 17%. The power for C_{max} of the

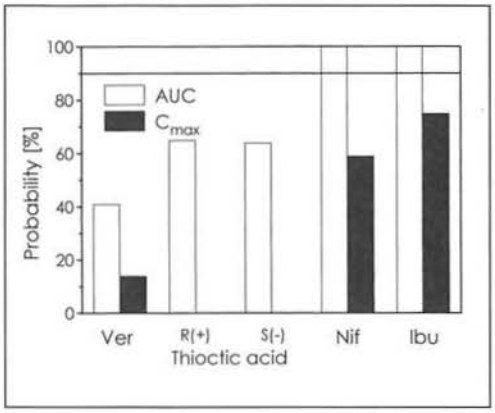


Figure 3: Probability for acceptance of bioequivalence of the reference product compared to itself following replicate single-dose administration of immediate release drug products containing verapamil, thioctic acid, nifedipin and ibuprofen. At least 16 subjects

thioctic acid enantiomers is even almost zero. Generally, the power could be increased by increasing number of subjects. This would, however, require the recruitment of several hundreds of subjects to demonstrate equivalence for C_{max} of $R(+)$ and $S(-)$ thioctic acid. These few examples show already that there are reference products whose bioequivalence compared to itself could only be demonstrated with low probability using the current criteria. In contrast, the generic product has to meet the criteria to receive approval.
These examples clearly underline the need for modification of the current approach of fixed bioequivalence limits. They suggest that the intraindividual variability of the reference product may be taken into account in order not to penalise the test formulation.

3 Population bioequivalence

In Figure 4 the outcome of a bioequivalence trial different to Figure 1 is displayed. Although showing a higher variability between the curves after administration of the test formulation compared to the reference product, the individual profiles result in similar mean curves. Evidently, the means are comparable but the variabilities are different. The higher variability following administration of the test formulation is reflected by a different population distribution. In contrast to average bioequivalence, test and reference product are defined to be bioequivalent according to population bioequivalence if the **entire POPULATION distributions** – particularly population means and variabilities – are "sufficiently similar" concerning AUC and C_{max}.
There have been proposed different metrics for statistical testing of population bioequivalence:

- Separate (disaggregate) metrics regarding

  mean $\mu_T - \mu_R$ and

  variance $\sigma_{TT}^2 - \sigma_{TR}^2$

  (BAUER and BAUER, 1994; HAUSCHKE, 1995)

σ_{TT}^2 and σ_{TR}^2 denote the total variance of AUC, C_{max} after administration of the test formulation and the reference product, respectively. The total variance results from between-subject and within-subject components.

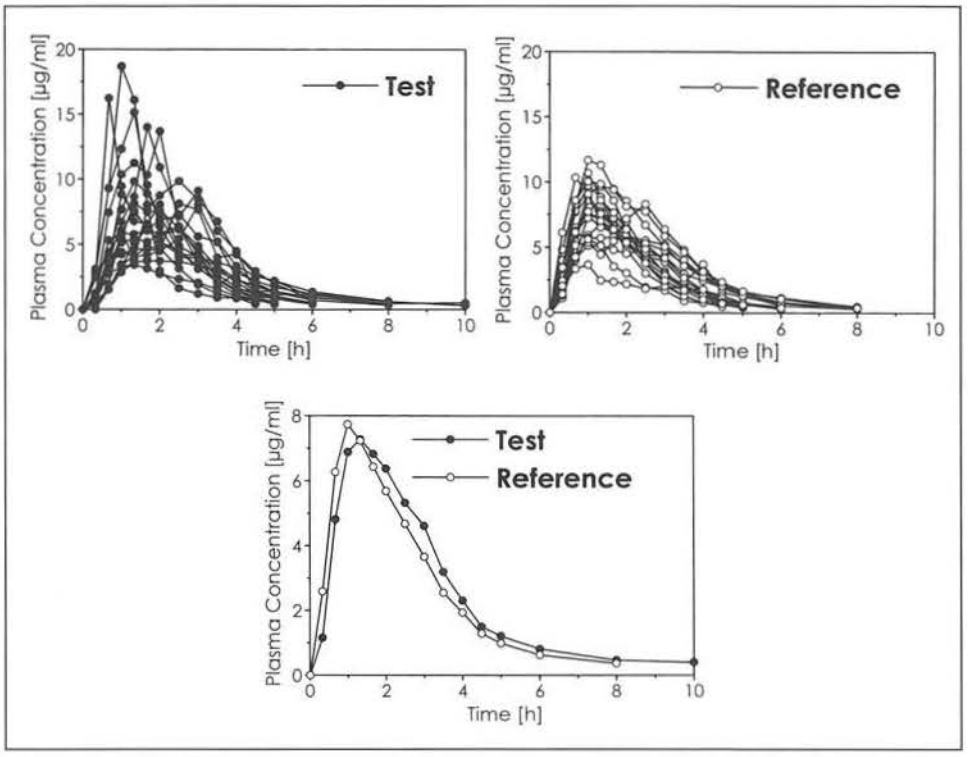


Figure 4: Individual and mean concentration vs. time profiles following administration of the test formulation in comparison to the reference product. 2-period crossover design. 18 subjects. Example for population bioequivalence. Different variabilities

- Aggregate metrics combining mean and variance e.g.

$$(\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2)$$

(SCHALL and LUUS, 1993)

- Scaled aggregate metrics combining mean and variance and scaling not to penalise the test formulation for high variability of the reference product e.g.

$$\frac{(\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2)}{\sigma_{\ddagger}^2},$$

where $\sigma_{\ddagger}^2 = \sigma_{TR}^2$ if $\sigma_{TR} > \sigma_{T0}$ reference-scaled

$\sigma_{\ddagger}^2 = \sigma_{T0}^2$ if $\sigma_{TR} \leq \sigma_{T0}$ constant-scaled

with σ_{T0}^2 constant related to a limit for the total variance

(FDA, 1997)

For statistical testing based on aggregate metrics there is no analytical solution available and bootstrap techniques have been proposed. For an overview of different approaches for constructing bootstrap confidence intervals for the assessment of bioequivalence see paper of PIGEOT (1999) in this issue.

4 Individual bioequivalence

The individual bioequivalence approach addresses switchability i.e. the individual subject's response (AUC, C_{max}) after administration of the test and reference product is expected to be similar. Thus, the test formulation in comparison to the reference product are considered to be bioequivalent if the **INDIVIDUAL subject means and variabilities** are "sufficiently similar" with regard to AUC and C_{max}. Average and population bioequivalence assume that the difference of individual subject's mean response under test compared to reference product is the same in each subject. But it may vary across subjects i.e. subject-by-formulation-interactions may occur.
In order to address switchability individual bioequivalence metrics take the following terms into account:

- Means: μ_T, μ_R
- Within-subject variances for T and R: σ^2_{WT}, σ^2_{WR}
- Subject-by-formulation-interaction variance component: σ^2_D

The metrics listed below were proposed to assess individual bioequivalence:

- Aggregate metrics combining mean and variances e.g.

  $$(\mu_T - \mu_R)^2 + \sigma^2_D + (\sigma^2_{WT} - \sigma^2_{WR})$$

  (SCHALL and LUUS, 1993)

- Scaled aggregate metrics combining mean and variances and scaling not to penalise the test formulation for high variability of the reference product e.g.

  $$\frac{(\mu_T - \mu_R)^2 + \sigma^2_D + (\sigma^2_{WT} - \sigma^2_{WR})}{\sigma^2_{\ddagger}},$$

  where $\sigma^2_{\ddagger} = \sigma^2_{WR}$ if $\sigma_{WR} > \sigma_{W0}$ reference-scaled

  $\sigma^2_{\ddagger} = \sigma^2_{W0}$ if $\sigma_{WR} \leq \sigma_{W0}$ constant-scaled

  with σ^2_{W0} constant related to a limit for the within-subject variance

  (FDA, 1997)

The limit for the within-subject variance could defined to be e.g. the limit for classifying highly variable drugs (30%).
Referring to the examples above, in the scaled aggregate metrics the intraindividual variability of the reference product is taken into account.
In analogy to the population bioequivalence, for statistical testing using aggregate metrics bootstrap methods also need to be applied. A description of the relevant bootstrap algorithms as well as a discussion of the statistical properties of bootstrap confidence intervals are given in the paper of PIGEOT (1999) in this issue. Practical experience in analysing data based on the individual bioequivalence approach is presented in the paper of PAPST (1999) in this issue.
Whereas the population bioequivalence approach is usually based on a 2-period crossover design with one administration of the test formulation and the reference product each, the individual bioequivalence criteria need a replicate changeover design. A 4-period, 2-sequence design (sequences: $T-R-T-R$ and $R-T-R-T$) and a 3-period, 2-sequence

design (sequences: $T-R-T$ and $R-T-R$) are recommended by the FDA (1997). These designs imply a higher level of requirements to the study conduct e.g. standardised study conditions over several periods. Volunteers will be confronted with increased stress. Additionally, a higher probability for drop-outs i.e. missing values in the analysis is expected.

5 Conclusions

- Population and individual bioequivalence concepts present solutions addressing drawbacks of the current practice of bioequivalence testing. They were developed in order to replace the average bioequivalence approach. Population and individual bioequivalence approaches are intended for different objectives of bioequivalence trials. Whereas population bioequivalence testing seems to be appropriate for preapproval changes of the investigational product during the process of new drug product development the individual bioequivalence approach is recommended in the context of registration of generic drug products and generic substitution as well as postapproval major changes (FDA, 1997).
- At present, population and individual bioequivalence are basically theoretical concepts which need empirical evidence and an elaboration of convincing. Moreover, aggregate metrics are artificial parameters which are difficult to interprete from a clinical and pharmaceutical as well as statistical point of view. Currently, the application of the new concepts is limited to a small number of bioequivalence trials published in the literature. Thus, practical experience is needed in the near future.
- Several metrics for assessing bioequivalence based on population and individual bioequivalence were created and could be used. Consequently, the question of how to choose the appropriate metric will be raised when planning a bioequivalence trial. For this purpose recommendations should be developed involving clinical, pharmaceutical and regulatory scientists.
- Aggregate metrics combining mean and variance as well as disaggregate metrics were proposed for assessing population bioequivalence. Whereas bioequivalence testing based on aggregate metrics allows a certain compensation of large differences in mean by small differences in variance (and vice versa), the use of disaggregate metrics provides separate results with regard to mean and variance. Further discussions among scientists should clarify the practical relevance of these two possibilities.
- For statistical testing of population and individual bioequivalence on the basis of aggregate metrics bootstrap techniques need to be applied. Bootstrapping is a statistical methodology which has not been widely applied and may be studied in routine.

References

BAUER, B., BAUER, M. M. 1994: Testing equivalence simultaneously for location and dispersion of two normally distributed populations. Biom. J. **36**, 643–660.

BLUME, H. H., MCGILVERAY, I. J., MIDHA, K. K. 1995a: Report of Consensus: Bio-International 94, Conference on bioavailability, bioequivalence and pharmacokinetic studies. Eur. J. Pharm. Sci. **3**, 113–124.

BLUME, H., MUTSCHLER, E. (eds.) 1995b: Bioäquivalenz – Qualitätsbewertung wirkstoffgleicher Fertigarzneimittel. Govi-Verlag, Eschborn, 5. Ergänzungslieferung.

CPMP WORKING PARTY ON EFFICACY OF MEDICINAL PRODUCTS, 1991: Note for Guidance III/54/89-EN: Investigation of bioavailability and bioequivalence.

FDA, DIVISION OF BIOEQUIVALENCE AND DIVISION OF BIOMETRICS, 1992: Statistical procedures for bioequivalence studies using a standard two-treatment crossover design.

FDA, CENTER FOR DRUG EVALUATION AND RESEARCH, 1997: Guidance for industry: In vivo bioequivalence studies based on population and individual bioequivalence approaches. Draft. Federal Register **62**, no. 249, Dec 30, 1997.

HAUSCHKE, D. 1995: A-priori ordered hypotheses in bioequivalence assessment. International Workshop: Statistical and regulatory issues on the assessment of bioequivalence, Düsseldorf.

MORAIS, J. A. G. 1995: Regulatory requirements for bioavailability studies of new active substances. In: Blume H. H., Midha, K. K. (eds.) Bio-International 2 Bioavailability, bioequivalence and pharmacokinetic studies, medpharm Scientific Publishers, Stuttgart, 171–179.

OHM, A. 1995: Critical manufacturing variables and in vitro dissolution tests in view of in vivo performance. In: Blume H. H., Midha, K. K. (eds.) Bio-International 2 Bioavailability, bioequivalence and pharmacokinetic studies, medpharm Scientific Publishers, Stuttgart, 261–279.

PAPST, G. 1999: Practical experiences with investigations of individual bioequivalence. Informatik, Biometrie und Epidemiologie in Medizin und Biologie **30**, 110–121.

PIGEOT, I. 1999: Comments on bootstrap confidence intervals for evaluating bioequivalence criteria. Informatik, Biometrie und Epidemiologie in Medizin und Biologie **30**, 96–109.

SCHALL, R., LUUS, H. G. 1993: On population and individual bioequivalence. Stat. Med. **12**, 1109–1124.

SCHUIRMANN, D. J. 1987: A comparison of the two-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J. Pharmacokin. Biopharm. **15**, 657–680.

Address for correspondence: Dr. Martina Elze, Pharmaceutical Research Associates, Besselstr. 2–4, D-68219 Mannheim, Germany. Tel: +49-621-8782 324, Fax: +49-621-8782 185, e-mail: ElzeMartina@pra-ww.com

Informatik, Biometrie und Epidemiologie in Medizin und Biologie **30** (3),
96–109, ISSN 0943-5581

Bootstrap Confidence Intervals for Evaluating Bioequivalence Criteria

Bootstrap-Konfidenzintervalle zur Erfassung von Bioäquivalenzkriterien

Iris Pigeot

Summary

In recent years, the concept of average bioequivalence has been criticised on certain respects which has led to the new concepts of individual and population bioequivalence. Based on these concepts new criteria for assessing bioequivalence have been proposed.
In this paper, different bioequivalence measures are reviewed from a statistical point of view. Due to the recommendations given in a draft of the FDA guidance (1997), we then focus on bootstrap confidence intervals for statistically evaluating the new bioequivalence criteria. For this purpose, the general idea of bootstrapping is repeated and different types of bootstrap intervals are presented. The percentile and the bias-corrected intervals are finally discussed in more detail following the proposals made by Schall and Luus (1993) and Schall (1995).

Keywords

BC_a method, bootstrap confidence intervals, individual bioequivalence, percentile method, population bioequivalence.

Zusammenfassung

In den letzten Jahren ist das klassische Konzept der Bioäquivalenz hinsichtlich verschiedener Aspekte kritisiert worden, was zu den neuen Konzepten der Populations- und der individuellen Bioäquivalenz geführt hat. Darauf aufbauend sind neue Kriterien zur Beurteilung der Bioäquivalenz vorgeschlagen worden.
In dieser Arbeit werden unter statistischen Gesichtspunkten verschiedene Bioäquivalenzmaße im Überblick dargestellt. Anschließend wird aufgrund der Empfehlungen in einem Entwurf der FDA-Richtlinien ein Schwerpunkt auf Bootstrap-Konfidenzintervalle zum statistischen Nachweis der Bioäquivalenz gelegt. Zu diesem Zweck werden zunächst die allgemeine Idee des Bootstraps wiederholt und verschiedene Ansätze zur Konstruktion von Bootstrap-Intervallen vorgestellt. Abschließend werden den Vorschlägen von Schall und Luus (1993) sowie von Schall (1995) folgend die sogenannten Perzentil- und bias-korrigierten Intervalle ausführlicher diskutiert.

Stichworte

BC_a-Methode, Bootstrap-Konfidenzintervalle, individuelle Bioäquivalenz, Perzentil-Methode, Populationsbioäquivalenz.

1 Introduction

The current approach for assessing bioequivalence of two different drug formulations is based on the concept of average bioequivalence. Since this concept only compares the mean bioavailability in two populations following from administration of a reference and a test formulation it has been more and more criticised and new concepts such as those of population and individual bioequivalence have been proposed (see among others ANDERSON, HAUCK, 1990). These new concepts now call for appropriate measures as for instance derived in ANDERSON and HAUCK (1990), SHEINER (1992), SCHALL and LUUS (1993), and SCHALL (1995). They can roughly be cross-classified into moment- and probability-based as well as in scaled and unscaled approaches (cf. SCHALL, WILLIAMS, 1996). Based on these measures criteria for assessing population and individual bioequivalence can be formulated. For the statistical evaluation of such criteria it is typically made use of confidence intervals. Here, you can find among others for the varying criteria quite a number of different approaches such as an exact confidence interval for evaluating average bioequivalence (LOCKE, 1984), a confidence interval based on an approximate F statistic (EKBOHM, MELANDER, 1989, ENDRENYI, 1994, 1995), maximum likelihood based methods (SHEINER, 1992), and bootstrap intervals (SCHALL, LUUS, 1993, SCHALL, 1995), where in this paper focus is on the latter method.
The paper is organised as follows. In the second section, moment- and probability-based criteria for assessing individual and population bioequivalence are presented. Some properties are pointed out. In Section 3, bootstrap techniques are introduced in general and different bootstrap intervals are discussed. The use of these bootstrap intervals for statistically evaluating the bioequivalence criteria is stressed in Section 4, where primarily the procedures and results of SCHALL and LUUS (1993) and SCHALL (1995) are reviewed. The paper closes with some final remarks.

2 Criteria for assessing bioequivalence

If one wishes to show that the bioavailabilities, characterised for instance by AUC (area under the concentration versus time curve) or $C_{\max}$ (observed maximum concentration), of two different formulations of the same substance, say reference and test formulation, are similar, bioequivalence studies and appropriate measures of bioequivalence are required. At present, the assessment of bioequivalence of two formulations is based on a comparison of the population means of their bioavailabilities and therefore typically called average bioequivalence. Average bioequivalence has been criticised on certain respects mainly by ANDERSON and HAUCK (1990). One objection concerns the fact that comparing the means only does not account for a possibly differing distribution of the bioavailabilities, not even for differing variances. Second, the fact that two formulations are bioequivalent with respect to their population means does not ensure that they are bioequivalent within one subject which is important under the aspect of switchability of formulations. Further drawbacks of the concept of average bioequivalence are addressed in the paper of ELZE (1999) in this issue.
Based on these concerns, two other approaches of bioequivalence are discussed in literature (cf. ANDERSON, HAUCK, 1990, SCHALL, LUUS, 1993): Population bioequivalence tends to assess bioequivalence by comparing the distributions of bioavailability, which is

of importance especially in case that the patient starts on a new drug, which is referred to as prescribability by HAUCK and ANDERSON (1992, 1994). In addition to the means, an appropriate measure for assessing population bioequivalence should at least compare the between-subject variances of the bioavailabilities.
To deal with the second objection, individual bioequivalence should be fulfilled, i.e. the responses to the reference and test formulation should not differ too much in the majority of patients. This implies that an appropriate measure should account for the within-subject variances. For a discussion of the different requirements for the assesssment of bioequivalence from a clinical and pharmaceutical perspective see ELZE (1999).
To be more precise, let us consider the following statistical models for the bioavailability Y_T from the test and Y_R from the reference formulation (cf. ANDERSON, HAUCK, 1990, SHEINER, 1992, SCHALL, LUUS, 1993, HOLDER, HSUAN, 1993):

$$\begin{aligned} Y_T &= \mu_T + b_T + \epsilon_T\,, \\ Y_R &= \mu_R + b_R + \epsilon_R\,, \end{aligned} \tag{2.1}$$

where μ_i, $i = T, R$, denotes the mean response in the ith population, b_i the mean deviation from the population average of a given individual, and ϵ_i the individual deviations from the individual mean response. The deviations b_i and ϵ_i are assumed as independent with $E(b_i) = E(\epsilon_i) = 0$ and with the within-subject variance $\mathrm{var}\,(\epsilon_i)$ denoted as $\sigma^2_{W_i}$ and the between-subject variance $\mathrm{var}\,(b_i)$ as $\sigma^2_{B_i}$. Thus, it follows

$$\begin{aligned} \mathrm{var}\,(Y_i) &= \mathrm{var}\,(b_i) + \mathrm{var}\,(\epsilon_i) \\ &= \sigma^2_{B_i} + \sigma^2_{W_i} = \sigma^2_i\,. \end{aligned} \tag{2.2}$$

For a given individual, b_R and b_T may be correlated where the variance of the within-subject difference of b_R and b_T is denoted as

$$\mathrm{var}\,(b_T - b_R) = \sigma^2_D. \tag{2.3}$$

It should be mentioned that the above additive models also apply in case that a multiplicative model is actually considered appropriate. Then, Y_i just represents the log-transformed data. Of course, it would be desirable to have a bioequivalence measure at hand which can be interpreted on the original scale.
Coming now to a more formal description of criteria for assessing the different approaches of bioequivalence given above, we first briefly look at the concept of average bioequivalence.
Two formulations are said to be bioequivalent with respect to their means, if the difference of μ_T and μ_R lies within a certain range, i.e.

$$-\Delta \leq \mu_T - \mu_R \leq \Delta \tag{2.4}$$

with Δ to be fixed by a drug regulatory authority. For log-transformed data let m_i, $i = T, R$, define the mean bioavailabilities of the two formulations on the original scale, then the formulations are regarded as bioequivalent if

$$\exp(-\Delta) \leq \frac{m_T}{m_R} \leq \exp(\Delta)\,. \tag{2.5}$$

The criteria for assessing population and individual bioequivalence to be presented here can be cross-classified as moment- or probability-based measures as well as in scaled and unscaled. Our presentation is essentially based on works of SCHALL and co-authors. For further details, we therefore refer to SCHALL and LUUS (1993), SCHALL (1995), and SCHALL and WILLIAMS (1996).

The main idea of the new criteria is to compare certain characteristics of the distribution of the difference between the bioavailabilities from the test and the reference formulation with the corresponding characteristics of the distribution of the difference between the two bioavailabilities from the reference formulation given twice. If the difference between test and reference formulation is not or only slightly greater than that within the reference formulation, the two formulations are said to be bioequivalent. Before going on, two questions should be stressed upon. First, how should the discrepancy between the bioavailabilities be measured? Second, do the differences between the bioavailabilities relate to between- or within-subject variations? The answer to the first question relates to the two proposals, i.e. moment- or probability-based, discussed in the following. The second question just concerns the concepts of population and individual bioequivalence.

In general, if we use the moment-based approach, two formulations are to be regarded as bioequivalent, if

$$E(Y_T - Y_R)^2 - E(Y_R - Y_R')^2 \leq \Delta_1^2, \tag{2.6}$$

where Y_T, Y_R, Y_R' denote the bioavailabilities following the administration of the test and of twice the reference formulation, Δ_1 has to be fixed by a drug regulatory authority. Please note, that the above equations (2.1), (2.2), (2.3), and (2.6) can be regarded as the basic underlying model. Now let

$$P_{TR} = P(|Y_T - Y_R| \leq r) \quad \text{and} \quad P_{RR} = P(|Y_R - Y_R'| \leq r), \tag{2.7}$$

then analogously, the probability-based approach considers two formulations as bioequivalent, if

$$P_{TR} - P_{RR} \geq \Delta_2. \tag{2.8}$$

In (2.7) and (2.8), r and Δ_2 have again to be fixed by drug regulatory authorities.

For assessing population bioequivalence, (2.6) has to be calculated for between-subject differences. Using (2.1) and the related assumptions and (2.2), we get

$$E(Y_T - Y_R)^2 = (\mu_T - \mu_R)^2 + \sigma_T^2 + \sigma_R^2, \tag{2.9}$$

$$E(Y_R - Y_R')^2 = \text{var}(Y_R - Y_R') = 2\sigma_R^2, \tag{2.10}$$

$$SD(Y_R - Y_R') = \sqrt{2}\sigma_R. \tag{2.11}$$

Thus, (2.6) results in

$$(\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2 \leq \Delta_1^2. \tag{2.12}$$

The probabilities in (2.8) have also to be calculated with respect to the between-subject variation, which are then additionally initialised with P to distinguish from the corresponding probabilities for assessing individual bioequivalence.

For assessing individual bioequivalence, (2.6) has to recalculated where now within-subject variations have to be accounted for. This yields

$$E(Y_T - Y_R)^2 = (\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{W_T}^2 + \sigma_{W_R}^2, \tag{2.13}$$

$$E(Y_R - Y_R')^2 = 2\sigma_{W_R}^2. \tag{2.14}$$

From (2.13) and (2.14), (2.6) results in

$$(\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{W_T}^2 - \sigma_{W_R}^2 \leq \Delta_1^2. \tag{2.15}$$

Calculating now (2.8) by accounting for the within-subject variation, the above probabilities are additionally indexed by I for individual bioequivalence.
Please note, that Δ_1^2 in (2.12) and in (2.15) as well as Δ_2 in (2.8) need not necessarily be fixed identical by the regulatory authority.
Taking a closer look to (2.12), it can easily be seen, that the population bioequivalence criterion reduces to that of average bioequivalence whenever the variances of the bioavailabilities from test and reference formulation are identical. This means that then the additional information on the variances does of course not lead to a "finer" criterion. A similar result can be obtained for (2.15). Let us first assume that the within-subject variabilities are identical under test and reference formulation. If now, in addition, $\sigma_D^2 = 0$, that is in total additional information on the behaviour of the bioavailability within subjects yield no information with respect to the above criterion, then (2.15) coincides with the average bioequivalence criterion provided that $\Delta_1 = \Delta$ in (2.4). Further relationships can be formulated for instance compared to the criterion by ANDERSON and HAUCK (1990) and among the above varying criteria (for details see SCHALL, LUUS, 1993, SCHALL, 1995).
The above moment- or probability-based criteria can also be used in scaled versions, i.e. scaled with respect to the variance of the reference formulation, which leads to a further distinction. A decision among these four criteria to assess bioequivalence should be made depending on the within-subject variability and the therapeutic range of a drug. A detailed discussion of how to proceed can be found in SCHALL and WILLIAMS (1996). The FDA guidance (draft from 1997) recommends to use a reference-scaled version, if $\sigma_R > \sigma_0$ or $\sigma_{W_R} > \sigma_{W_0}$, where σ_0^2 and $\sigma_{W_0}^2$ are the variance and within-subject variance of the reference formulation, respectively, from which point on the scaled criterion has a wider bioequivalence range compared to its unscaled counterpart (see also SCHALL, WILLIAMS, 1996). The FDA recommends to use 0.2 as value for the above unknown standard deviations, whereas SCHALL and WILLIAMS (1996) suggested that for instance σ_{W_0} should be chosen smaller than 0.3.

3 The bootstrap

For convenience, let us now first recall the idea of resampling techniques and especially that of the bootstrap. We refer here in particular to the book by EFRON and TIBSHIRANI (1993) and especially regarding a detailed discussion of the asymptotic behaviour of bootstrap confidence intervals to the book by SHAO and TU (1995, Chapter 4).
Resampling techniques are very computer-intensive methods for statistical inference. Fast and powerful computers necessary for their effective use are now widely available, which makes these techniques easily accessible to practitioners in their daily work. Roughly speaking, resampling techniques are based on a repeated use of parts of a data set. They are often applied for non-parametrically estimating measures of accuracy for statistical estimates as for instance bias, variance, confidence intervals or even for estimating the entire distribution of an estimator. As a special representative of such methods we focus here on the bootstrap, which is known as a data-based simulation technique and which has been introduced by EFRON (1979). The main idea of the bootstrap is described in the following very intuitive manner by looking at two worlds: the real world and the bootstrap world (EFRON and TIBSHIRANI, 1993, p. 87; cf. Figure 3.1). In the real world we want to get information on an unknown probability distribution F where we typically focus on certain parameters $\theta(F)$. To get this information we draw a sample $\mathbf{x} = (x_1, \ldots, x_n)$ from which we calculate a statistic of interest $\hat{\theta} = g(\mathbf{x})$. To assess the accuracy of this estimator $\hat{\theta}$ of θ we would like to know, for instance, its bias $\text{bias}_F(\hat{\theta}, \theta)$ or its variance $\text{var}_F(\hat{\theta})$, which are, however, unknown since they are calcu-

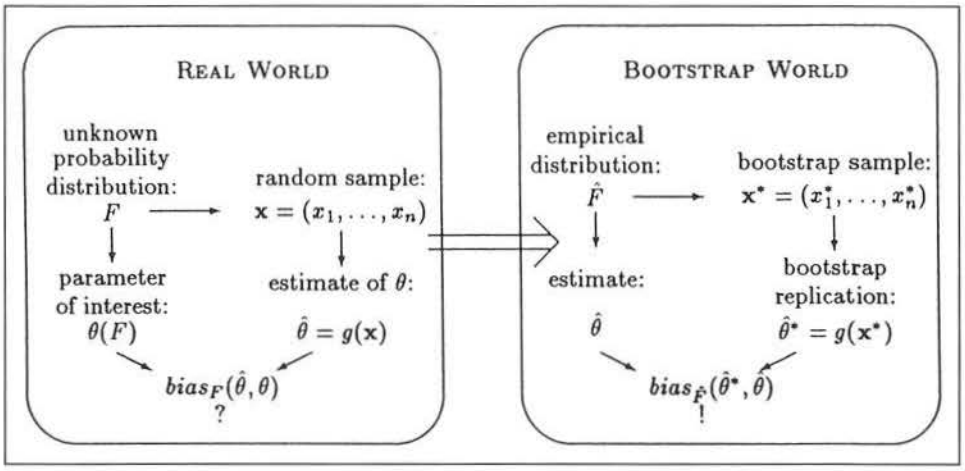


Figure 3.1: Schematic diagram of the bootstrap

lated with respect to the unknown probability distribution F. In the bootstrap world, the unknown probability distribution F is replaced by some reasonable estimator $\hat{F}$ as, for instance, the empirical distribution function. Now, an arbitrary number of bootstrap samples $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$ can be drawn from $\hat{F}$. From each of these samples we calculate the estimate θ of interest, denoted as bootstrap replications $\hat{\theta}^* = g(\mathbf{x}^*)$. Thus, all is known in this artificial world which enables us to now calculate for instance the bias of $\hat{\theta}^*$ compared of course to $\hat{\theta}$ with respect to $\hat{F}$, i.e. $\text{bias}_{\hat{F}}(\hat{\theta}^*, \hat{\theta})$. This idea is at the first glance very appealing, although it is at the same time the crucial point, since all bootstrap calculations are based on the replacement of F with $\hat{F}$. Thus, all properties of bootstrap estimators regarding the real world depend on the behaviour of $\hat{F}$ as estimator of F.

This brings us back to the original concept of the bootstrap as plug-in estimator, which will be illustrated in the following for the case that the variance of $\hat{\theta}$ is to be estimated. As already mentioned above, the popular version to obtain a bootstrap estimate of $\text{var}_F(\hat{\theta})$ is via a Monte-Carlo algorithm which consists of the following steps. First, an estimator $\hat{F}$ of F has to be determined. In the second step, bootstrap samples of size n are generated from $\hat{F}$ with $X_1^*, \ldots, X_n^*$ i.i.d. from $\hat{F}$ given $X_1, \ldots, X_n$ from which $\hat{\theta}^* = \hat{\theta}(X_1^*, \ldots, X_n^*)$ is calculated. Typically, $\hat{F}$ is chosen as the empirical distribution function $\hat{F}_n$ which means that the bootstrap samples can be drawn with replacement from the population $x_1, \ldots, x_n$. In the next step, K independent bootstrap samples are drawn, which yield K bootstrap replications $\theta_1^*, \ldots, \theta_K^*$, from which we get an approximate bootstrap variance estimator as the empirical variance of the bootstrap replications

$$\widehat{\text{var}}_{\text{boot}}^{\text{appr}}(\hat{\theta}) = \frac{1}{K} \sum_{k=1}^{K} \{\hat{\theta}_k^* - \bar{\hat{\theta}}_{(\cdot)}^*\}^2$$

with $\bar{\hat{\theta}}_{(\cdot)}^* = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k^*$.

However, this estimator is only an approximation of the exact bootstrap variance estimator because of the finite number of replications. For deriving an exact bootstrap variance estimator, consider the general definition of the variance of an estimator $\hat{\theta}$ which reads

as follows

$$\begin{aligned}\mathrm{var}_F(\hat{\theta}) &= E_F(\hat{\theta} - E_F(\hat{\theta}))^2 \\ &= \int (\hat{\theta}(X_1, \ldots, X_n) - \int \hat{\theta}(X_1, \ldots, X_n)\, dF(x)^n)^2\, dF(x)^n \end{aligned} \tag{3.1}$$

with $X_1, \ldots, X_n$ i.i.d. from F, $x \in R$. If the unknown distribution F in (3.1) is replaced with some appropriate estimator $\hat{F}$ we get the bootstrap variance estimator, i.e.

$$\widehat{\mathrm{var}}_{\mathrm{boot}}(\hat{\theta}) = \int (\hat{\theta}(X_1, \ldots, X_n) - \int \hat{\theta}(X_1, \ldots, X_n)\, d\hat{F}(x)^n)^2\, d\hat{F}(x)^n . \tag{3.2}$$

This estimator can be regarded as the conditional variance of $\hat{\theta}(X_1^*, \ldots, X_n^*)$ given the original sample $X_1, \ldots, X_n$ denoted as

$$\widehat{\mathrm{var}}_{\mathrm{boot}}(\hat{\theta}) = \mathrm{var}_*(\hat{\theta}(X_1^*, \ldots, X_n^*) \mid X_1, \ldots, X_n), \tag{3.3}$$

where the bootstrap sample $X_1^*, \ldots, X_n^*$ is i.i.d. from the estimated distribution function $\hat{F}$.

Typically, the formulae (3.1) and (3.2), however, cannot be written as explicit functions of F or $\hat{F}$, respectively. Then, $\widehat{\mathrm{var}}_{\mathrm{boot}}(\hat{\theta})$ can only be calculated approximately, where the algorithm described above is the most popular approximation.

As already mentioned, the bootstrap is a flexible method with a broad range of possible applications. Thus, it can among others be used to estimate the sampling distribution H_F of $\hat{\theta}$ with

$$H_F(x) = P\{\hat{\theta}(X_1, \ldots, X_n;\ F) \le x\} \tag{3.4}$$

by

$$\hat{H}_{\mathrm{boot}}(x) = P_*\{\hat{\theta}(X_1^*, \ldots, X_n^*;\ \hat{F}_n) \le x \mid X_1, \ldots, X_n\}, \tag{3.5}$$

where $P_*\{\cdot \mid X_1, \ldots, X_n\}$ means the conditional probability of $\hat{\theta}(X_1^*, \ldots, X_n^*)$ given $X_1, \ldots, X_n$. (3.5) can be approximated using the above algorithm by

$$\hat{H}_{\mathrm{boot}}^{\mathrm{appr}}(x) = \frac{1}{K} \sum_k I\{\hat{\theta}(X_{1k}^*, \ldots, X_{nk}^*;\ \hat{F}_n) \le x\}, \tag{3.6}$$

where $I(\cdot)$ denotes the indicator function.

The bootstrap distribution estimator is consistent for many widely used estimators, but it depends on the tail behaviour of F such that its consistency requires more stringent moment conditions than those typically required for an existing limit distribution of $\hat{\theta}_n$. In addition, a certain degree of smoothness of $\hat{\theta}$ is necessary for $\hat{H}_{\mathrm{boot}}$ being consistent.

One of the most common applications of the bootstrap is the calculation of confidence intervals. As a first approach, one can use the bootstrap to estimate quantiles of the distribution G of the studentized pivot

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}} \quad \text{with} \quad \hat{\sigma} = \sqrt{\widehat{\mathrm{var}}_F(\hat{\theta})}$$

and to then calculate a bootstrap-t confidence interval (EFRON, 1982). The distribution G can be estimated as

$$\hat{G}_{\mathrm{boot}}(x) = P_* \left\{ \frac{\hat{\theta}_{\mathrm{boot}} - \hat{\theta}}{\hat{\sigma}_{\mathrm{boot}}} \le x \mid X_1, \ldots, X_n \right\}, \tag{3.7}$$

where $\hat{\theta}_{\mathrm{boot}}$ and $\hat{\sigma}_{\mathrm{boot}}$ are bootstrap analogs of $\hat{\theta}$ and $\hat{\sigma}$. The confidence bounds are then obtained as

$$\hat{\theta} - \hat{\sigma}\, \hat{G}_{\mathrm{boot}}^{-1}\left(1 - \frac{\alpha}{2}\right) \quad \text{and} \quad \hat{\theta} - \hat{\sigma}\, \hat{G}_{\mathrm{boot}}^{-1}\left(\frac{\alpha}{2}\right) \tag{3.8}$$

to achieve a $(1-\alpha)$ confidence interval. This confidence interval lacks from the fact that a variance estimator $\hat{\sigma}^2$ is needed. In the case that no simple estimator $\hat{\sigma}$ is available, a nested bootstrap is possibly required. In addition, this interval is not invariant under reparametrisations due to the studentized pivot. However, using the studentized pivot may be necessary for calculating "traditional" confidence intervals where knowledge of the distribution of the pivot is needed to determine the quantiles, but constructing a pivot is no longer necessary with the bootstrap. This implies that confidence intervals for an unknown parameter θ can be directly derived from the bootstrap distribution of an estimator $\hat{\theta}$ of θ without knowing the sampling distribution of $\hat{\theta}$ or of its pivot. Thus, EFRON (1981) suggested the bootstrap percentile confidence interval

$$\left[\hat{H}_{\text{boot}}^{-1}\left(\frac{\alpha}{2}\right);\ \hat{H}_{\text{boot}}^{-1}\left(1-\frac{\alpha}{2}\right)\right], \tag{3.9}$$

where $\hat{H}_{\text{boot}}^{-1}\left(\frac{\alpha}{2}\right)$ denotes the $100 \cdot \frac{\alpha}{2}$th percentile of

$$\hat{H}_{\text{boot}}(x) = P_*\{\hat{\theta}_{\text{boot}} \leq x \,|\, X_1, \ldots, X_n\}.$$

These percentiles can be approximated by the $100 \cdot \frac{\alpha}{2}$th empirical percentile $\hat{\theta}_K^*\left(\frac{\alpha}{2}\right)$ of the bootstrap replications $\hat{\theta}_k^*$, $k = 1, \ldots, K$, which means the $K \cdot \frac{\alpha}{2}$th value in the ordered list of the K bootstrap replications. This results in the following approximate interval

$$\left[\hat{\theta}_K^*\left(\frac{\alpha}{2}\right);\ \hat{\theta}_K^*\left(1-\frac{\alpha}{2}\right)\right]. \tag{3.10}$$

It may be appropriate to accommodate for a bias term and for a possible skewness when calculating a confidence interval for $\hat{\theta}$ (cf. EFRON, 1982, 1987). This results in the so-called bootstrap accelerated bias-corrected percentile, briefly denoted as BC_a, where an acceleration constant a is introduced to measure the rate of change of the standard deviation of $\hat{\theta}$ with respect to the true parameter value θ. The BC_a interval is given as

$$\left[\hat{H}_{\text{boot}}^{-1}(\alpha_1);\ \hat{H}_{\text{boot}}^{-1}(\alpha_2)\right] \tag{3.11}$$

with

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})}\right), \tag{3.12}$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})}\right), \tag{3.13}$$

where Φ denotes the cdf of the standard normal distribution and $\Phi(z_{\alpha/2}) = \frac{\alpha}{2}$. Note that α_1 equals $\Phi(z_{\alpha/2}) = \alpha/2$ and $\alpha_2 = \Phi(z_{1-\alpha/2}) = 1 - \alpha/2$ in case that $\hat{z}_0 = 0$ and $\hat{a} = 0$, which means that the BC_a interval coincides with the simple percentile interval in case that no bias and no acceleration have to be accounted for. EFRON and TIBSHIRANI (1993, p. 186) suggested to compute $\hat{z}_0$ as

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}_k^* < \hat{\theta}\}}{K}\right), \tag{3.14}$$

where $\hat{z}_0$ can roughly be interpreted as the median bias of $\hat{\theta}_{\text{boot}}$ in normal units.

As easiest way to compute $\hat{a}$ EFRON and TIBSHIRANI proposed to use the jackknife. Let $\hat{\theta}_{-i} = \hat{\theta}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ and $\bar{\hat{\theta}}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{-i}$, then $\hat{a}$ can be obtained as

$$\hat{a} = \frac{\sum_{i=1}^{n} (\bar{\hat{\theta}}_{(\cdot)} - \hat{\theta}_{-i})^3}{6 \left\{ \sum_{i=1}^{n} (\bar{\hat{\theta}}_{(\cdot)} - \hat{\theta}_{-i})^2 \right\}^{3/2}} . \tag{3.15}$$

An advantage of this interval is that it is transformation respecting, which means that the BC_a bounds transform correctly if instead of θ a function of it is considered. Actually, the BC_a automatically chooses its best scale (EFRON, TIBSHIRANI, 1993, p. 187).
Of course, it is of interest to know the performance of the intervals presented above, especially concerning their coverage probabilities. It can be shown (SHAO, TU, 1995, p. 142) that the coverage probabilities of the above intervals all converge to the nominal levels for n tending to infinity. This is fulfilled for the bootstrap-t, if $\hat{G}_{\text{boot}}$ is consistent for the sampling distribution of the studentized pivot. For the percentile interval and the BC_a, the above property holds if $\hat{H}_{\text{boot}}$ is consistent and the sampling distribution of $\hat{\theta}$ shows a certain asymptotic behaviour. A second desirable property of confidence intervals concerns its accuracy, where a confidence interval CI_n of θ is said to be lth order accurate if

$$P\{\theta \in CI_n\} = 1 - \alpha + O(n^{-l/2}) . \tag{3.16}$$

The bootstrap-t and the equal-tailed two-sided percentile and BC_a are all second-order accurate, whereas the one-sided percentile is only first-order accurate (cf. SHAO, TU, 1995, p. 144ff).

4 Bootstrap intervals for bioequivalence measures

In this section, different approaches for constructing bootstrap intervals for statistically evaluating individual or population bioequivalence are reviewed according to the proposals by SCHALL and LUUS (1993) and SCHALL (1995). The bioequivalence measures are estimated from cross-over studies. For a discussion of various cross-over designs regarding their impact on the performance of the resulting confidence intervals see PABST (1999) in this issue.

4.1 A bootstrap percentile interval

SCHALL and LUUS (1993) derived bootstrap percentile intervals for moment- as well as probability-based measures of individual and population bioequivalence. Their proposals are here only given for situations without a period effect.

Population bioequivalence

As introduced in Section 2, Equations (2.12) and (2.8), we consider the following bioequivalence measures

$$\theta_1 = (\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2 , \tag{4.1}$$

$$\theta_2 = P_{TRP} - P_{RRP} , \tag{4.2}$$

which can be estimated from a typical two-period cross-over study. The data obtained from such studies with in total n subjects can be written as $(y_T, y_R) = [(y_{T1}, y_{R1}), \ldots, (y_{Tn}, y_{Rn})]$, where n_1 subjects receive the treatment sequence $T - R$ and the remaining $n_2 = n - n_1$ subjects the sequence $R - T$. Based on these data an unbiased estimator of θ_1 reads as

$$\hat{\theta}_1 = (\hat{\mu}_T - \hat{\mu}_R)^2 - \hat{\sigma}_d^2/n + \hat{\sigma}_T^2 - \hat{\sigma}_R^2, \tag{4.3}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ denote the sample means and the sample variances of y_i, $i = T, R$, and $\hat{\sigma}_d^2$ the sample variance of $(y_T - y_R)$. An unbiased estimator of the probability-based measure θ_2 can be obtained as difference of the relative frequencies

$$\hat{P}_{TRP} = \frac{\#\Big((y_{Ti}, y_{Rj}),\, i \neq j;\, |y_{Ti} - y_{Rj}| \leq r\Big)}{n(n-1)}, \tag{4.4}$$

$$\hat{P}_{RRP} = \frac{\#\Big((y_{Ri}, y_{Rj}),\, i < j;\, |y_{Ri} - y_{Rj}| \leq r\Big)}{n(n-1)/2} \tag{4.5}$$

as estimators of the unknown probabilities P_{TRP} and P_{RRP}. Using now the percentile method (Equation (3.10)) for calculating confidence intervals for θ_1 and θ_2 based on the above estimators yields the following algorithm:

Step 1: Draw a bootstrap sample (y_T^*, y_R^*) of size n with replacement from the original sample (y_T, y_R).

Step 2: Calculate $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$ from the bootstrap sample.

Step 3: Repeat Step 1 and 2 K times with K typically chosen as 1000. Then, the $(1 - \alpha)$ quantile of the bootstrap replications $\hat{\theta}_{11}^*, \ldots, \hat{\theta}_{1K}^*$ gives the upper bound of a one-sided $(1 - \alpha)$ confidence interval for θ_1. Analogously, we obtain the lower bound of a one-sided $(1 - \alpha)$ confidence interval for θ_2 as the α quantile of the bootstrap replications $\hat{\theta}_{21}^*, \ldots, \hat{\theta}_{2K}^*$.

Individual bioequivalence

For assessing the individual bioequivalence we use

$$\theta_1 = (\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{W_T}^2 - \sigma_{W_R}^2, \tag{4.6}$$

$$\theta_2 = P_{TRI} - P_{RRI} \tag{4.7}$$

(cf. Equations (2.15), (2.8)), which can for instance be estimated using data $(y_T, y_{R1}, y_{R2}) = [(y_{T1}, y_{R11}, y_{R21}), \ldots, (y_{Tn}, y_{R1n}, y_{R2n})]$ from a three-period cross-over study where the test formulation is given once and the reference formulation twice to each subject of a sample of size n. Since $\theta_1 = E(Y_T - Y_R)^2 - E(Y_R - Y_R')^2$ (cf. (2.6)), an unbiased estimator of θ_1 can be obtained as difference of unbiased estimators of each of these expectations with

$$\hat{E}(Y_T - Y_R)^2 = \frac{\sum_{i=1}^{n} [(y_{Ti} - y_{R1i})^2 + (y_{Ti} - y_{R2i})^2]}{2n}, \tag{4.8}$$

$$\hat{E}(Y_R - Y_R')^2 = \frac{\sum_{i=1}^{n} (y_{R1i} - y_{R2i})^2}{n}. \tag{4.9}$$

Analogously to the corresponding measure of the population bioequivalence, an unbiased estimator of the probability-based measure θ_2 can be derived from the relative frequencies

$$\hat{P}_{TRI} = \frac{1}{2n}\left\{\#\left((y_{Ti}, y_{R1i});\ |y_{Ti} - y_{R1i}| \leq r\right) + \#\left((y_{Ti}, y_{R2i});\ |y_{Ti} - y_{R2i}| \leq r\right)\right\}, \tag{4.10}$$

$$\hat{P}_{RRI} = \frac{\#\left((y_{R1i}, y_{R2i});\ |y_{R1i} - y_{R2i}| \leq r\right)}{n} \tag{4.11}$$

as estimators of the unknown probabilities P_{TRI} and P_{RRI}. Using again the percentile method for calculating confidence intervals for θ_1 and θ_2 we get the following algorithm:

Step 1: Draw a bootstrap sample $(y_T^*, y_{R1}^*, y_{R2}^*)$ of size n with replacement from the original sample (y_T, y_{R1}, y_{R2}).

Step 2: Calculate $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$ from the bootstrap sample.

Step 3: Repeat Step 1 and 2 K times with K typically chosen as 1000. Then, the $(1-\alpha)$ quantile of the bootstrap replications $\hat{\theta}_{11}^*, \ldots, \hat{\theta}_{1K}^*$ gives the upper bound of a one-sided $(1-\alpha)$ confidence interval for θ_1. Analogously, we obtain the lower bound of a one-sided $(1-\alpha)$ confidence interval for θ_2 as the α quantile of the bootstrap replications $\hat{\theta}_{21}^*, \ldots, \hat{\theta}_{2K}^*$.

Note that the above algorithms are essentially the same, but the estimates are of course calculated with respect to the concept of bioequivalence being of interest, which means that the between- or within-subject variation has to be considered.
In the presence of a period effect the above algorithms have to be modified such that the treatment sequence is accounted for. This means among others that for instance in the algorithms for calculating confidence intervals for the population bioequivalence measures a sample of n_1 pairs has to be drawn from those subjects with sequence $T-R$ and a sample of n_2 pairs from those with sequence $R-T$ (for details see SCHALL, LUUS, 1993, p. 1117f).

4.2 A bias-corrected bootstrap interval

In contrast to the proposal of SCHALL and LUUS (1993), SCHALL (1995) derived a bias-corrected bootstrap interval for probability-based measures of individual and population bioequivalence. Thus, he considered again (2.7), but now using a modified version of these probabilities where r is no longer fixed but formulated as a multiple γ of the standard deviation of $Y_R - Y_R'$. This standard deviation has to be calculated with respect to the between- or within-subject variation depending on whether individual or population bioeqivalence should be assessed. The resulting bioequivalence criteria can be based on the difference or on the ratio of the above probabilities. For a discussion of the choice of γ we refer to SCHALL (1995).

Population bioequivalence

Being interested in the population bioeqivalence (2.7) then results in

$$P_{TRP} = P(|Y_T - Y_R| \leq \gamma\sqrt{2}\,\sigma_R)$$

and P_{RRP} defined analogously.

SCHALL then distinguished two situations depending on the assumptions concerning the distribution of the bioavailabilities. If no further assumptions are made he proposed a distribution-free approach comparable to that of Section 4.1. Here, an unbiased estimator of $SD(Y_R - Y_R') = \sqrt{2}\,\sigma_R$ is obtained from a two-period cross-over as

$$\sqrt{2}\,\hat{\sigma}_R = \left[\frac{1}{n(n-1)/2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (Y_{Ri} - Y_{Rj})^2 \right]^{1/2},$$

from which we then get estimators of P_{TRP} and P_{RRP} as in (4.4) and (4.5), but with r chosen as $\gamma\sqrt{2}\,\hat{\sigma}_R$. Then, the algorithm for calculating a bias-corrected bootstrap interval according to (3.11) with $\hat{a} = 0$ since no acceleration is accounted for is given as:

Step 1: Calculate $\hat{P}_{TRP}$ and $\hat{P}_{RRP}$ from the original data.
Step 2: Draw a bootstrap sample (y_T^*, y_R^*) of size n with replacement from the original sample (y_T, y_R).
Step 3: Calculate $\hat{P}_{TRP}^*$ and $\hat{P}_{RRP}^*$ as well as their difference from the bootstrap sample.
Step 4: Repeat Step 2 and 3 K times with K typically chosen as 1000.
Step 5: Calculate $z_\alpha = \Phi^{-1}(\alpha)$ and $\hat{z}_0$ as

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{P}_{TRP,k}^* - \hat{P}_{RRP,k}^* < \hat{P}_{TRP} - \hat{P}_{RRP}\}}{K} \right).$$

Then, the $\Phi(z_\alpha + 2\hat{z}_0)$ quantile of the bootstrap replications gives the lower bound of a one-sided $(1-\alpha)$ confidence interval for $P_{TRP} - P_{RRP}$.

Please note that no period effect has been accounted for in the above algorithm. An analogous algorithm for the ratio of P_{TRP} and P_{RRP} is readily obtained.

Let us now assume that the bioavailabilities Y_T and Y_R are normally distributed, such that the probability P_{TRP} can be expressed in terms of the standard normal distribution function as

$$P_{TRP} = \Phi\left(\frac{\gamma\sqrt{2}\,\sigma_R + \mu_T - \mu_R}{\sqrt{\sigma_R^2 + \sigma_T^2}} \right) - \Phi\left(\frac{-\gamma\sqrt{2}\,\sigma_R + \mu_T - \mu_R}{\sqrt{\sigma_R^2 + \sigma_T^2}} \right)$$

(SCHALL, 1995), whereas P_{RRP} is a constant. Then, the above algorithm reads as follows:

Step 1: Calculate $\hat{P}_{TRP}$ from an ANOVA of the original data.
Step 2: Draw a bootstrap sample (y_T^*, y_R^*) of size n with replacement from the original sample (y_T, y_R).
Step 3: Calculate $\hat{P}_{TRP}^*$ from an ANOVA of the bootstrap sample.
Step 4: Repeat Step 2 and 3 K times with K typically chosen as 1000.
Step 5: Calculate $z_\alpha = \Phi^{-1}(\alpha)$ and $\hat{z}_0$ as

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{P}_{TRP,k}^* < \hat{P}_{TRP}\}}{K} \right).$$

Then, the $\Phi(z_\alpha + 2\hat{z}_0)$ quantile of the bootstrap replications gives the lower bound of a one-sided $(1-\alpha)$ confidence interval for P_{TRP}.

Note that in case of a present period effect an ANOVA of the data has to be carried out with subject, formulation, and period as main effects. In addition, Step 2 of the above algorithm has to be done in a stratified manner, i.e. the bootstrap observations must be drawn from those original observations with the same treatment sequence.

Individual bioequivalence

Bias-corrected bootstrap intervals for evaluating individual bioequivalence measured by the above probabilities, now denoted as P_{TRI} and P_{RRI}, can be derived analogously. That is in the two algorithms described above P_{TRP} and P_{RRP} and their estimators have just to

be replaced with P_{TRI} and P_{RRI}, where, of course, the estimators of the latter probabilities have to be calculated from at least a three-period cross-over study. Thus, the bootstrap samples have to be drawn from at least triples of observations. If no normal distribution is assumed for the bioavailabilities, P_{TRI} and P_{RRI} have again to be estimated via the corresponding relative frequencies (cf. (4.10) and (4.11)). If we assume normally distributed bioavailabilities, P_{TRI} can be calculated as

$$P_{TRI} = \Phi\left(\frac{\gamma\sqrt{2}\,\sigma_{W_R} + \mu_T - \mu_R}{\sqrt{\sigma^2_{W_R} + \sigma^2_{W_T} + \sigma^2_D}}\right) - \Phi\left(\frac{-\gamma\sqrt{2}\,\sigma_{W_R} + \mu_T - \mu_R}{\sqrt{\sigma^2_{W_R} + \sigma^2_{W_T} + \sigma^2_D}}\right)$$

and P_{RRI} again is a constant.
In SCHALL (1995) some simulation results are also reported which indicate that the bias-corrected intervals in general show very good coverage probabilities. For further details see SCHALL (1995).

4.3 FDA Guidance

In the preceding sections, various proposals for constructing bootstrap intervals based on different bioequivalence measures have been presented. In a draft dated 1997, the FDA now recommends the use of moment-based measures. The according criteria might be reference- or constant-scaled depending on the estimated value of the total variance σ^2_R of the reference formulation when assessing population bioequivalence or on the estimated within-subject variance $\sigma^2_{W_R}$ of the reference formulation in case of individual bioequivalence, respectively.
Following the proposal of SCHALL and LUUS (1993), an appropriate 90% equal-tailed confidence interval derived from the bootstrap percentile method is suggested for statistically evaluating the bioequivalence criteria. The idea behind a two-sided equal-tailed 90% confidence interval is that the upper confidence bound relates to a statistical test with the significance level of 5% (cf. FDA draft, 1997). In addition, the FDA draft recommends the percentile bootstrap interval and not the BC_a as appropriate method because of current experience. The interval should be approximated from at least 2000 bootstrap samples in contrast to the usual number of 1000 replicates where the treatment sequence should be preserved. That means bootstrapping has to be done in a stratified manner to account for a possible period effect as mentioned at the end of Section 4.1.
Regarding the choice of a reference- or constant-scaled version of the criteria it should be mentioned in this context that this decision is only to be based on the original data set. That is, once the decision has been made, this selection should be used for all bootstrap samples regardless of the specific estimates of σ^2_R or $\sigma^2_{W_R}$ obtained from each of the bootstrap samples.
Further details on the choice of the study design and the statistical models can be taken from that draft.

5 Discussion

The focus of this paper was on discussing statistical aspects of the new bioequivalence criteria and not the appropriateness of the reviewed bioequivalence measures. These proposals are certainly to be reconsidered as it is for instance done in some of the other papers on bioequivalence in this issue. Here, other approaches can be thought of perhaps based on other study designs.
But even if the introduced concepts are considered as appropriate, the properties of the bootstrap intervals still remain as an open question which are worth to be investigated in

more detail. This means on the one hand a theoretical discussion of the asymptotic properties of the theoretical bootstrap intervals for the presented bioequivalence measures including their asymptotic coverage probabilities and their accuracy. On the other hand, the stability of the bootstrap algorithms should be further examined in terms of reproducibility and exactness since decisions on the approval of a formulation will be made based on the calculated intervals. One possibility for avoiding a possible allegation of a misuse would be to specify starting values for the bootstrap algorithms in the study protocol in advance. Besides this aspect, it is of course still of interest to judge the behaviour of the confidence intervals in finite-sample situations not only regarding their coverage probabilities, but also in terms of power.

References

ANDERSON, S., HAUCK, W. W. (1990): Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* **18**, 259–273.

EFRON, B. (1979): Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **7**, 1–26.

EFRON, B. (1981): Nonparametric standard errors and confidence intervals (with discussion). *Canadian Journal of Statistics* **9**, 139–172.

EFRON, B. (1982): *The Jackknife, the Bootstrap, and Other Resampling Plans*. SIAM, Philadelphia.

EFRON, B. (1987): Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association* **82**, 171–200.

EFRON, B., TIBSHIRANI, R. J. (1993): *An Introduction to the Bootstrap*. Chapman & Hall, New York.

EKBOHM, G., MELANDER, H. (1989): The subject-by-formulation interaction as a criterion of interchangeability of drugs. *Biometrics* **45**, 1249–1254.

ELZE, M. (1999): Bioequivalence trials – status and perspectives. To appear in: *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* **3**, 87–95.

ENDRENYI, L. (1994): A method for the evaluation of individual bioequivalence. *International Journal of Clinical Pharmacology and Therapeutics* **32**, 497–508.

ENDRENYI, L. (1995): A simple approach for the evaluation of individual bioequivalence. *Drug Information Journal* **29**, 847–855.

FDA (1997): In Vivo Bioequivalence Studies Based on Population and Individual Bioequivalence Approaches. Guidance for Industry, Center for Drug Evaluation and Research, Food and Drug Administration (Draft, date: October 1997).

HAUCK, W. W., ANDERSON, S. (1992): Types of bioequivalence and related statistical considerations. *International Journal of Clinical Pharmacology and Therapeutics* **30**, 181–187.

HAUCK, W. W., ANDERSON, S. (1994): Measuring switchability and prescribability: when is average bioequivalence sufficient? *Journal of Pharmacokinetics and Biopharmaceutics* **22**, 551–564.

HOLDER, D. J., HSUAN, F. (1993): Moment-based criteria for determining bioequivalence. *Biometrika* **80**, 835–846.

LOCKE, C. S. (1984): An exact confidence interval from untransformed data for the ratio of two formulation means. *Journal of Pharmacokinetics and Biopharmaceutics* **12**, 649–655.

PABST, G. (1999): Practical experience with investigations of individual bioequivalence. To appear in: *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* **3**, 110–121.

SCHALL, R. (1995): Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar. *Biometrics* **51**, 615–626.

SCHALL, R., LUUS, H. G. (1993): On population and individual bioequivalence. *Statistics in Medicine* **12**, 1109–1124.

SCHALL, R., WILLIAMS, R. L. (1996): Towards a practical strategy for assessing individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* **24**, 133–149.

SHAO, J., TU, D. (1995): *The Jackknife and Bootstrap*. Springer-Verlag, New York.

SHEINER, L. B. (1992): Bioequivalence revisited. *Statistics in Medicine* **11**, 1777–1788.

Address for correspondence: Prof. Dr. Iris Pigeot, Institute of Statistics, University of Munich, Ludwigstraße 33, 80539 München, Germany. Tel.: +49-89-2180 3522, Fax: +49-89-2180 5041, e-mail: pigeot@stat.uni-muenchen.de

Informatik, Biometrie und Epidemiologie in Medizin und Biologie **30** (3),
110–121, ISSN 0943-5581

Practical Experiences with Investigations of Individual Bioequivalence

Praktische Erfahrungen mit Untersuchungen zur individuellen Bioäquivalenz

G. Pabst

Summary

Individual bioequivalence, at least when using the procedures proposed in a recent draft guidance of the FDA, can only be assessed in a repetitive design trial. This paper in the first part, based on preceding experience, summarizes aspects that should be taken into account during set-up and performance of repetitive design trials. In the second part, the statistical procedures proposed by the FDA are applied to actual data sets.
The data indicate that repetitive designs with 3 periods should be avoided because of their lower power and since the statistical methods are applicable only with great difficulties. A repetitive design with 4 periods on the other hand might require too large a total blood volume to be withdrawn. Furthermore, the design of the data actually available for evaluation will not be balanced in most cases. The cross-over design with multiple dosing in two periods where concentration-time profiles are derived on two consecutive days should be investigated more intensively as a potential alternative.
The proposed methods for the statistical evaluation of intraindividual bioequivalence on the one hand are biased and on the other hand as iterative procedures much too frequently lead to results of limited reliability at the border of the parameter space or even do not converge at all. The methodology of evaluation needs to be investigated more intensively before an actual application in clinical trials can be recommended.

Key words

Individual bioequivalence, repetitive design, SAS PROC MIXED, bootstrap, applicability

Zusammenfassung

Individuelle Bioäquivalenz kann nur in Studien mit repetitivem Design untersucht werden, zumindest wenn man die in einer Draft Guidance der FDA vorgeschlagenen Methoden verwendet. Die vorliegende Publikation diskutiert im ersten Teil, basierend auf vorliegenden Erfahrungen, die Aspekte, die bei Planung und Durchführung von Studien im repetitiven Design zu beachten sind. Im zweiten Teil werden die von der FDA vorgeschlagenen statistischen Methoden auf Datensätze aus vorausgegangenen Studien angewandt.

Die vorliegenden Erfahrungen und Ergebnisse zeigen, daß repetitive Designs mit 3 Perioden vermieden werden sollten wegen ihrer geringeren Trennschärfe aber auch weil die statistischen Methoden nur mit großen Schwierigkeiten angewandt werden können. Bei einem repetitiven Design mit 4 Perioden kann anderseits das zu entnehmende Blutvolumen unakzeptabel groß werden. Zudem wird das Design der tatsächlich zur Auswertung zur Verfügung stehenden Daten in der Regel nicht mehr balanciert sein. Das cross-over Design mit multipler Dosierung in zwei Perioden, bei dem Konzentrations-Zeit-Profile an zwei aufeinanderfolgenden Tagen untersucht werden, könnte eine gangbare Alternative sein.
Die vorgeschlagenen Methoden zur statistischen Auswertung der individuellen Bioäquivalenz sind einerseits verzerrt (biased) und führen anderseits als iterative Verfahren viel zu häufig zu Ergebnissen eingeschränkter Zuverlässigkeit am Rand des Parameterraums, sofern sie überhaupt konvergieren. Die Auswertungsmethodik muß intensiver untersucht werden, bevor ein Einsatz in tatsächlichen klinischen Studien empfohlen werden kann.

Schlüsselwörter

Individuelle Bioäquivalenz, repetitive Designs, SAS PROC MIXED, Bootstrap, Anwendbarkeit

1 Introduction

The FDA has issued a draft guidance (for comment purposes only, not yet for implementation) indicating an intention to changed requirements away from average bioequivalence to population and individual bioequivalence (FDA, 1997). For a description of the different concepts of bioequivalence and the rationale behind them, see e.g. the paper of ELZE (1999) in this issue. Although the theoretical basis for these approaches is fairly well established, practical experience with actual clinical trials is lacking.
Individual bioequivalence, at least when using the procedures proposed by the FDA, can only be assessed in a repetitive design where each subject receives at least one of the treatments more than once. The separate occasions of treatment administration customarily are called "periods" and the sequence of administration is designated e.g. by (TRTR), denoting an application of the test treatment T in the first, of the reference treatment R in the second period, of T in period 3 and R in period 4.
This communication in the first part summarizes aspects that should be taken into account during set-up and performance of repetitive design trials. In the second part, the statistical procedures proposed by the FDA are applied to actual data sets in order to get a first impression on their applicability in a real-life situation.
The paper attempts to address almost all aspects at least cursorily. The customary structure "introduction – material and methods – results – discussion" therefore was applied to individual sections but not to the paper as a whole. Further practicability aspects of the proposed methods for individual bioequivalence are discussed by ENDRENYI et al. (1998).

2 Data sets available

Using the procedures proposed by the FDA, individual bioequivalence can only be assessed in a repetitive design. On the other hand, a standard non-repetitive cross-over is sufficient for assessment of average bioequivalence, which is the current requirement. In consequence, only very few repetitive pharmacokinetic trials have been performed in the

past. In order to allow the industry to gain some experience, the FDA has placed data of 12 repetitive trials on their internet site (FDA, 1998a), 10 trials with 4 periods and 2 trials with 3 periods. In some of them up to 6 substances were investigated, however noting that in case of an endogenous substance pharmacokinetic results with and without baseline correction were posted as separate data sets. Two of these data sets were identified as stemming from trials performed in the clinical facilities of the company the author is affiliated with, and three additional repetitive trials on file (one 4-period and two 3-period trials) also will be considered when discussing the implications of a repetitive design on study performance. On the other hand, the sparse data in the literature, see e.g. GRAHNÉN et al. (1984), ESINHART and CHINCHILLI (1994) or SHUMAKER and METZLER (1998), will be disregarded because of a conceivable publication bias and because most of the trials reported upon were not designed as repetitive trials from the onset.

Altogether there are 11 trials in the data base with 4 periods. None of them used the design recommended by the FDA with two sequences (TRTR) and (RTRT). The data base, however, includes two trials performed according to a two-sequence design using treatment sequences (TRRT) and (RTTR). One trial was performed with 6 different sequences of treatment order, all other 4-period trials ($n = 8$) involved four different sequences. There are only 4 trials in the data base with a 3-period design. None of them was performed according to the design proposed by the FDA using only two treatment sequences (TRT) and (RTR).

All of the trials were performed in order to assess average bioequivalence. However, as far as performance aspects are concerned, all these trials nevertheless can be and were taken as practical experience. On the other hand, the practicability of the proposed statistical evaluation method was assessed for the data on the FDA internet site exclusively.

3 Points to consider in set-up and performance of repetitive design trials

3.1 Design

The FDA recommends to use designs with two sequences only and with a cross-over between consecutive periods, leading to the four-period design with sequences (TRTR) and (RTRT) and, as a possible alternative, to a three-period design with sequences (TRT) and (RTR), however noting that "a greater number of subjects would be needed for the three-period design compared to the recommended four-period design to achieve the same statistical power" (FDA, 1997). The power of repetitive designs for assessment of average bioavailability is treated by CHEN et al. (1997).

The designs proposed by the FDA are not balanced for first-order carry-over, since the test treatment T always is followed by the reference treatment R but never by itself and vice versa. Such designs are suitable only in case that carry-over is of no concern. The FDA guidance, on the other hand, discusses the implications of carry-over and sequence effects quite extensively.

Intraindividual bioequivalence is investigated as a surrogate of switchability of patients between drug products during an ongoing therapy. However, the relevance of a single dose trial for an investigation of switchability has to be questioned. A multiple dose design, where the pharmacokinetics are investigated on two consecutive days at steady-state might be much more relevant. This design is mentioned by the FDA in a sidemark as a possible alternative. If, according to this design, the same preparation is given on two consecutive days, the within-treatment variability is a measure of true variability from day to day. If, alternatively, the preparation is switched from one day to the next, results of the first day might be used to assess differences in steady-state levels attained

and results on the next day will mimic any problems that might occur upon switching. This design, whether with or without switching from one steady-state day to the next, in our experience is much easier to perform and in addition offers the advantage that results of the first steady-state day may be evaluated according to the standard methods of average bioequivalence. A more intensive discussion of this design, however, would exceed the scope of this paper.

3.2 Subjects and clinical performance

Repetitive design trials require participation of the subjects for a longer duration than in a standard two-period cross-over. Recruitment thus is somewhat more difficult and the time requirements might in effect exclude the working population from participation, although the draft guidance stipulates that "restrictions to entry into the study should be based solely on safety considerations".

The longer duration of a repetitive trial increases the risk of subjects dropping-out due to personal reasons or for intermittent adverse events that might not even have anything to do with the trial, e.g. taking on a job in another city. In fact, among the 11 trials with 4 periods there was only a single one where the design of the actually available data, take C_{max} for example, was balanced in the sense that all treatment sequences were represented in the data set equally often – although it may be assumed that they were designed to being balanced. Only one of the 4 trials with 3 periods was balanced.

A replacement of drop-out subjects will result in a considerable prolongation of the overall study duration. There is the temptation to include replacements only if the trial otherwise would be inconclusive. Thus, there is a risk of managers asking for an (unplanned) interim evaluation which, however, has considerable statistical implications. The FDA therefore recommends to include additional subjects in order to be prepared for drop-outs. In view of the facts cited above, this "surplus" should be larger than in a standard two-period cross-over, but even then there is no guarantee that the final sample size available for evaluation will be at least as high as the anticipated number. If some of the extra subjects in fact should not be needed because of a lower drop-out rate than anticipated, the FDA allows that their samples will not be analyzed if the study protocol contains a statement to this effect. According to this approach, some subjects might be dosed and blood might be withdrawn from them without being actually used later-on. The ethical justification of such a procedure may be questioned.

The repetitive design with 3 periods has several disadvantages as will be shown later-on. On the other hand, a repetitive design with 4 periods might not be possible because of too high a blood draw. For example, 4 periods with 15 samples of 10 mL each and 3 laboratory exams requiring 45 mL each would sum up to a total blood withdrawal of 735 mL. Ways out of this dilemma would be in a more refined analytical method, requiring a lower sample volume, or in fewer samples being withdrawn, thus, however, leading to concentration-time profiles that are described with less detail.

3.3 Analytical aspects

Customarily all of the samples of a single subject are analyzed within the same analytical run. With the higher total number of samples being withdrawn this might not be possible anymore, thus introducing an additional unavoidable variability of results. Lastly, because of the longer duration of the trial in each subject, the stability of analytical samples over a sufficiently long period of time has to be confirmed in course of the validation of the analytical method.

3.4 Pharmacokinetic aspects

The area under the concentration-time curve AUC is the pharmacokinetic target parameter used for assessment of the extent of bioavailability, respectively extent of exposure. Based on the general relation $\mathrm{AUC} = \frac{f \cdot \mathrm{D}}{\mathrm{CL}}$, an intraindividual comparison of this parameter is considered to reflect differences in the absolute bioavailability f, based on the assumption that the clearance CL is an intraindividual constant. However, it is not clear, whether this assumption can be upheld for the longer duration of a repetitive trial. If CL should shift, this will introduce some additional variability. The same restrictions also apply to $C_{\max}$, since this pharmacokinetic parameter is strongly correlated with AUC; in the most simple one-compartment model it may be expressed as AUC times a function depending only on absorption and elimination rate constants.

4 The statistical evaluation

4.1 Descriptive statistics

Some thought should be given on how to display results in a reasonable manner. Individual listings and graphs of concentration-time profiles should allow to distinguish the profiles taken for the same treatment in different periods. Difficulties arise when it comes to plotting mean concentration-time profiles. It is not reasonable to calculate means (and standard deviations) over all individual data sets derived for each treatment, since subjects contribute more than one observation. SHUMAKER and METZLER (1998) for example report results of a repetitive 4-period design trial using means and standard deviations calculated separately for first and second replication of test and reference treatment, respectively, thus leading to four mean curves. Another conceivable, globally applicable approach would be to perform statistical analyses at each time point (ANOVA using GLM methodology) in order to derive least-squares means with the associated standard errors. This method, in our opinion, is to be favored because it results in two mean curves only.

4.2 Test statistics for individual bioequivalence

The test statistic to be evaluated is

$$\frac{(\mu_{\mathrm{T}} - \mu_{\mathrm{R}})^2 + \sigma_{\mathrm{D}}^2 + (\sigma_{\mathrm{WT}}^2 - \sigma_{\mathrm{WR}}^2)}{\max(\sigma_{\mathrm{WR}}^2,\ \sigma_{\mathrm{W0}}^2)} \quad \text{with} \quad \sigma_{\mathrm{W0}}^2 = 0.2 \tag{4.1}$$

for which it has to be shown that it is lower than a critical value θ_I.
μ_{T} and μ_{R} stand for the global means of the pharmacokinetic parameter of interest of test (T) and reference treatment R, σ_{WT}^2 and σ_{WR}^2 are the within-subject variances for T and R and σ_{D}^2 is the subject-by-formulation interaction variance component. The criterion was cited here exactly as specified in the draft guidance of the FDA (1997), which also provides a derivation of the formulae. The relation to other and alternative bioequivalence criteria is discussed e.g. by LIU and CHOW (1997) and PIGEOT (1999). The actual values of θ_I and σ_{W0}^2 to be used are fixed by the authorities and are of no concern for the present paper dealing with the practical applicability of the proposed methodology.
This test statistic was developed for and shall be applied to data after logarithmic pretransformation exclusively. The first objective therefore is to estimate the terms in the

test statistic, i.e. $(\mu_T - \mu_R)$, the difference between test and reference mean of the logarithms of the pharmacokinetic parameter under investigation, σ^2_{WT} and σ^2_{WR}, the within-subject (intraindividual) variability under test and reference treatment, respectively, and σ^2_D, the subject-by-formulation interaction. Estimates for $(\mu_T - \mu_R)$, σ^2_{WT} and σ^2_{WR} can be derived by mixed effects modeling and σ^2_D may be estimated based on the between-subject (interindividual) variabilities σ^2_{BT} and σ^2_{BR} of test and reference treatment, respectively, and the within-subject correlation ϱ between individual treatment means, considering that

$$\sigma^2_D = (\sigma_{BT} - \sigma_{BR})^2 + 2(1 - \varrho)\,\sigma_{BT}\sigma_{BR}\,. \tag{4.2}$$

Inspection of the formula reveals that σ^2_D can be zero only if there is a perfect correlation ($\varrho = 1$) and only if $\sigma_{BT} = \sigma_{BR}$, whereas under actual circumstances the estimates for σ^2_{BT} and σ^2_{BR}, i.e. $\hat{\sigma}^2_{BT}$ and $\hat{\sigma}^2_{BR}$, always will differ simply due to sampling variability. σ^2_D therefore generally will be overestimated. ENDRENYI and TOTHFALUSI (1999) simulated data sets with $\sigma^2_D = 0$, but this subject-by-formulation interaction nevertheless was estimated to be larger than 0.15 in a fairly large number of trials – the limit of 0.15 is used by the FDA as a criterion to assess a possible clinical relevance of the interaction. Finally, it should be taken into account when interpreting results, that σ^2_{WT} and σ^2_{WR} are positively correlated.

The parameters may be estimated based on statistical moments, see CHINCHILLI (1996) and CHINCHILLI and ESINHART (1996). This procedure, however, presently is not endorsed yet by the FDA but in future may very well be the procedure of choice. The procedure favored by the FDA in the guidance is based on mixed effects modeling using restricted maximum likelihood estimates (REML), see FDA (1998b).

4.3 SAS code for evaluation of the mixed effects model

The internet site of the FDA exhibits SAS® statements that might be used for mixed effects modeling, refer to Table 1, which now shall be discussed in some detail.

Table 1: SAS statements recommended by the FDA for evaluation

```
PROC MIXED DATA=d METHOD=REML MAXITER=200
    NOITPRINT NOCLPRINT ORDER=FORMATTED;
CLASS treat subj period sequ;
MODEL ln&var = treat period sequ period*sequ(treat);
RANDOM treat / SUBJECT=subj TYPE=CSH;
REPEATED / GROUP=treat;
LSMEANS treat / DIFF=CONTROL;
ESTIMATE 'T - R' treat -1 1;
RUN;
```

Statements written in italics are optional; the period by sequence interaction period*sequ(treat) only needs to be evaluated when not using one of the recommended designs with two sequence groups.

Let us now address some of the features of the evaluation:

4.3.1 Number of iterations

MAXITER=200: The default in SAS is 50 iterations. For the data sets on the FDA internet site, always considering AUC = AUC(0–∞) and C_{max}, PROC MIXED some-

times did not converge for the 3-period designs, but for the 4-period designs convergence always was reached within 3 to 8 iterations (median: 6) and the number of iterations needed did not exhibit any systematic trend whether the model included the period *sequ(treat) term or not.

4.3.2 Internal order of factor levels

ORDER=FORMATTED: SAS by default sorts fixed effect factors, i.e. those in the CLASS statement, in alphabetical order. This should be kept in mind when specifying the factor loading in the ESTIMATE statement or the CONTRAST statement which also may be used. If for example the test preparation is labeled as '1' and the reference as '2', then the factors should be 1 and −1, in this order. The ordering however has even higher implications for the option DIFF=CONTROL where always the first code in the order specified is used as reference. Lastly, the output shows the interindividual between-subject variances as Var(1) and Var(2) in the implicit order, however without specifying which is which. In alphabetical order, using codes 'R' and 'T', the estimate of σ^2_{BR} is shown as Var(1) and that of σ^2_{BT} as Var(2). This immediately leads to the standard output.

Table 2: Sample output of the evaluation using SAS (excerpt with annotation of variance estimates)

DATA SET pard17A − ln (C_{max})

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Group	Estimate	
Var(1)	SUBJ		0.47493127	σ_{BR}^2
Var(2)	SUBJ		0.46279981	σ_{BT}^2
CSH	SUBJ		0.94789176	ϱ
DIAG		TREAT R	0.12333438	σ_{WR}^2
DIAG		TREAT T	0.17102135	σ_{WT}^2

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
TREAT	1	70	2.11	0.1511
PERIOD	3	70	1.81	0.1525
SEQU	1	70	1.63	0.2054
PERIOD*SEQU(TREAT)	2	70	0.22	0.8018

ESTIMATE Statement Results

Parameter	Estimate	Std Error	DF	t	Pr > |t|
T − R	−0.10571213	0.07283063	70	−1.45	0.1511

Least Squares Means

Effect	TREAT	LSMEAN	Std Error	DF	t	Pr > |t|
TREAT	R	5.57918141	0.12047100	70	46.31	0.0001
TREAT	T	5.47346927	0.12177863	70	44.95	0.0001

Differences of Least Squares Means

Effect	TREAT	_TREAT	Difference	Std Error	DF	t	Pr > |t|
TREAT	T	R	−0.10571213	0.07283063	70	−1.45	0.1511

4.3.3 SAS Output

Table 2 shows an excerpt of a typical output produced by SAS using the statements of Table 1; an evaluation of ln (C_{max}) of data set 17 A posted on the internet site of the FDA.
The results of the ESTIMATE statement and those labeled "Differences of least squares means", which are the results of the DIFF=CONTROL option, should coincide. This may be used as a reliability check. The estimates shown are estimates for ($\mu_T - \mu_R$).

4.3.4 Estimability

In 20 of the 40 tests performed on 4-period designs (data sets of the FDA, always considering AUC and C_{max}, with and without the period*sequ(treat) interaction term) the SAS procedure led to the Note: Estimated G matrix is not positive definite. Such a situation theoretically should have been avoided by the use of REML estimates and according to the SAS documentation "this is usually not a cause for concern". For the tests performed here, the note commonly was associated with an estimated within-subject correlation $\hat{\varrho}$ (listed as CSH SUBJ) of 1.0 – which result, although theoretically possible, never will be true in an actual data situation. Too high an estimate for ϱ will lead to a lower estimate of σ_D^2. This however even might be of advantage, considering that σ_D^2 generally tends to be overestimated. There was no relevant difference in the frequency of a G matrix not positive definite, independent whether the statistical model included the period*sequ(treat) interaction term or not. Furthermore, there was no obvious dependency on the pattern of missing data. In contrast to expectations, for 10 of the 40 analyses the G matrix was not positive definite for ln (C_{max}) whereas the evaluation of ln (AUC(0−∞)) of the same data set, which parameter here and then could not be estimated and thus is missing, led to a positive definite G matrix.
The SAS procedure PROC MIXED could be applied to data of the partially repetitive designs with 3 periods only with difficulty. Whenever the model included the period *sequ(treat) interaction, least squares treatment means and their difference never could be estimated using the standard SAS statements as proposed – estimates however were always available for the model excluding the interaction term. The FDA in fact explicitly notes that "for three-period designs, the simple estimate statement is usually not enough. The coefficients of the estimable function need to be specified". Correct specification of the estimable function however is quite tricky, if possible at all. Furthermore, since it depends on the internal ordering of the data, the SAS statement by itself might not be sufficient to document that the factors specified actually were correct in order to estimate what they purport to do.
At last, it should be mentioned that the initial parameter estimates automatically generated by SAS in some of the 3-period designs were not feasible "Note: Initial estimate is not feasible". This problem could be overcome by explicitly specifying initial estimates, e.g. using the statement PARMS (0.4) (0.4) (0.8) (0.2) (0.2); Under such circumstances, however it should be shown separately that the final estimates are not dependent on the initial values selected.
By the way, since PROC MIXED is an iterative routine, it never can be assured that this procedure will converge at the absolute minimum or only at a local one, if it converges at all. Some assurance might be gained if the procedure were recalculated with different sets of initial values (besides those automatically determined by SAS), hoping that it will converge to the same parameter set (within a reasonable numerical precision). Use of the internal SAS routine for derivation of start values for the parameter estimates at least makes the procedure reproducible.

4.4 Period effects

The statistical evaluation always extracts period effects, although the impact of period effects, if there should be any, generally is not discussed. For a two-period design they are only of little concern, since only subjects who contribute data of both periods can reasonably by used during the evaluation anyhow. For repetitive 3- and 4-period designs, especially if there are drop-outs, a period effect might, at minimum, reduce the reliability of the results derived.

4.5 Normality

The procedure implicitly assumes that variances correspond to a normal distribution. For the standard two-period designs this may be tested easily, e.g. by considering the residuals in one of the two periods. For the repetitive designs evaluated here, this approach is not possible, since the residuals are correlated. Lastly, even if the normality is under doubt, corresponding nonparametric approaches are not available. The procedure furthermore is only stated for ratios, respectively data after logarithmic pre-transformation. A corresponding procedure suitable for an evaluation on the original untransformed scale is not available, although for example the half-value duration HVD and t_{max} should be evaluated on the arithmetic scale, if at all. Furthermore, the possible values that t_{max} may take on are pre-determined by the sampling times and therefore should not be assumed to be normally distributed.

4.6 Bootstrap confidence intervals

The significance of the proposed test statistic cannot be assessed by comparing it to some theoretical distribution. The FDA therefore recommends to calculate 95% upper confidence bounds for the test statistic by bootstrapping and to conclude individual bioequivalence if the upper end of the CI is lower than the pre-fixed margin θ_I. Bootstrapping should use a minimum of 1500 (2000 recommended) bootstrap samples. The evaluation of the original data determines whether σ^2_{WR} or σ^2_{W0} shall be used in the denominator of (4.1) for the all the bootstrap samples. The FDA recommends use of the percentile method without mentioning acceleration or bias-correction. For an overview of the different types of bootstrap confidence intervals for the assessment of bioequivalence see the paper of PIGEOT (1999) in this issue.
In the following, a brief description of the resampling procedure and its realization in SAS is given.

4.6.1 Resampling procedure

Bootstrapping shall be performed stratified by sequence group and always selecting all observations of a subject simultaneously. This manner of proceeding should assure that the structure of subjects within sequence is retained. However, as noted above, for 10 of the 11 trials performed in a 4-period design, the data sets included one or more subjects with incomplete data. Bootstrapping whole subjects propagates the structure of missing values and even may enhance it. For example, if a sequence group should consist of only six subjects, a specific subject theoretically will never be selected for 33.5% of the bootstrap samples, will be selected once in 40.2% of cases, twice in 20.1%, 3 times in 5.4%, and 4 to 6 times in 0.9% of cases. If the subject should be an extreme outlier, the 95% bootstrap confidence interval will correspond to a bootstrap sample where the subject had been selected three times. If the subject should have been a drop-out contribut-

ing only data of two periods, the confidence interval similarly might relate to a bootstrap sample where this subject was represented three times among the six bootstrap subjects in the sequence group. In most cases this however will not be of major concern, since the FDA favors designs with two sequence groups only, such that the sample sizes in the sequence groups will be larger.
When programming the bootstrap resampling procedure, care should be taken that artificial subject numbers are assigned in order to distinguish sets of observations in case that subjects were selected more than once for the specific bootstrap sample. For example, if subject no. 3 would be selected a second time, this subject's data might be duplicated assigning subject no. 103, the third time using subject no. 203 etc.

4.6.2 Realization in SAS

The FDA strongly advises "Do not use a BY statement to bootstrap PROC MIXED" although there is no theoretical reason behind this advice. However, if generating 2000 bootstrap samples of a trial with 24 subjects in a 4-period design, this would result in 192000 observations and the data set just may become too unwieldy to handle. Furthermore, since evaluation of 2000 bootstrap sets may last several hours, it is wise not to use BY-processing in order that interim results are saved even if the computer system should break down at some time in-between. Lastly, there is the concern that the start values of the parameters in PROC MIXED might be influenced by results in the preceding BY group – the SAS documentation does not contain any statement to such a dependence or lack thereof. The recommended procedure therefore is the use of MACRO programming.
SAS for Windows includes a new 'Output Delivery System' for PROC MIXED which might be used to restructure the standard output. The main advantage, however, is that all the output may be routed to SAS data sets while suppressing the standard output, e.g. by the statement MAKE 'ESTIMATE' OUT=e NOPRINT;. Information in these data sets then may be used for saving and collecting the relevant results of each bootstrap case. However, if PROC MIXED should not converge or if the initial estimate was not feasible, the corresponding data set will not be generated! Furthermore, some but not all of the data sets will be missing if PROC MIXED should lead to "Final Hessian is not positive definite". In order to anticipate such cases, the MACRO should first generate dummies with missing results for all of the data sets to be generated by PROC MIXED which will be replaced by actual data if everything went well.
Unfortunately the standard comments and notes cannot be routed to data sets. A way around this is to use PROC PRINTTO to route the LOG output to some file which then may be checked in a separate DATA step for notes.

4.6.3 Plausibility and non-convergence of PROC MIXED

When combining bootstrap results, they should be checked for plausibility. For example, it might happen that the within-subject correlation ϱ is estimated to be negative. Furthermore, a G matrix not positive definite generally will lead to estimates at the border of the parameter space. Before proceeding to the bootstrap confidence interval it therefore might be reasonable first to inspect results for e.g. $|\mathsf{Var(2)}| = \hat{\sigma}^2_{\mathrm{BT}} < 10^{-8}$ or $\mathsf{CSH} = \hat{\varrho} > 1 - 10^{-8}$. This is a real problem especially for the three-period designs, however also for 4-period designs. Bootstrapping of ln (AUC) of data set 8 of the FDA, a 4-period design with 19 subjects in 4 sequence groups, had the following results:

0.7% not evaluable, e.g. because of an infeasible initial estimate or no convergence
7.4% final Hessian not positive definite
0.6% ϱ estimated to be negative

15.9% estimated G matrix not positive definite, always leading to $1 - \hat{\varrho} < 10^{-8}$
1.6% with selection of a single subject in one of the sequence groups (this should not be a problem if one of the two-sequence designs is used, as recommended)

Thus, in total there were more than 8% bootstrap samples with missing results and 18% of questionable merit. Even if only using the 92% of bootstrap samples with results, it is clear that this puts severe limits on the reliability of the confidence interval. By theoretical reasons, and in order to be on the safe side, any bootstrap samples with missing results preferably should be assumed to represent the most extreme cases. With more than 5% missing results it thus will not be possible to give a (one-sided) 95% confidence interval.
However, even for a seemingly well-behaved data set, e.g. ln (C_{max}) of set 17 A with 40 subjects in 4 sequence groups, the G matrix was not positive definite in 25.2% of all bootstrap cases.

4.6.4 Randomness

Selection of bootstrap samples will depend on the initial value of the random number generator. This might be specified explicitly, but commonly the random number generator is initialized using the internal clock time of the processor. Two different calculation runs therefore have to be expected to lead to slightly different results, although the bootstrap CI commonly will have stabilized after about 500 iterations – and the FDA recommends to draw at least 1500, better 2000 bootstrap samples. However, because of the conceivable differences, there is the temptation to rerun the bootstrap in case of borderline results. In order to protect against this, the value to be used for initializing the random number generator might be specified in the study protocol or the statistical analysis plan, or it might be specified that the upper end of the confidence interval will be rounded to perhaps 3 significant digits before comparing it with θ_I.

4.6.5 Processing time

More of a practical than of a theoretical concern is the overall longer time needed for an evaluation. Whereas an evaluation of average bioequivalence might be completed in perhaps 10 minutes per parameter, calculation of the bootstrap confidence interval (2000 samples) will take several hours: In our experience, SAS for Windows, Win 95 in a local Novell network, about 2.5 hours if calculation runs in the foreground, up to 7–8 hours as a background process.

5 Summary

Evaluation of intraindividual bioequivalence necessitates the use of repetitive designs, preferably one with 4 periods. The most relevant restriction on the applicability of repetitive designs is the larger total blood volume to be taken from each subject, in some cases precluding such a design. The 3-period design on the other hand has only a lower discriminatory power and, most importantly, the proposed statistical procedure can be applied to it only under quite severe restrictions.
In contrast to a two-period design, there is a high risk that some subjects will provide only incomplete data. The final design of the data available for evaluation therefore in most cases will be unbalanced. The procedures to be applied should be able to handle such data correctly and the impact of this on the reliability of overall results should be discussed.

Whereas most of the problems can be overcome, at least for the four-period design with two sequences, the most relevant argument against the proposed procedure, from a point of practical applicability, is the fact that PROC MIXED sometimes does not lead to any results at all and in about ¼ of cases leads to results of questionable merit (G matrix not positive definite with parameter estimates at the margin of the parameter space). Furthermore, the proposed method for the statistical evaluation of intraindividual bioequivalence tends to over-estimate the subject-by-formulation interaction. More investigation into the methods is needed before an application can be recommended in actual clinical trials.

References

CHEN, K. W., CHOW, S. C., LI, G. (1997): A note on sample size determination for bioequivalence studies with higher-order crossover designs. J. Pharmacokin. Biopharm. **25**, 753–765.

CHINCHILLI, V. M. 1996: The assessment of individual and population bioequivalence. J. Biopharm. Stat. **6**, 1–14.

CHINCHILLI, V. M., ESINHART, J. D. (1996): Design and analysis of intra-subject variability in crossover experiments. Stats. in Med. **15**, 1619–1634.

ELZE, M., BLUME, H. H. (1999): Bioequivalence trials – Status and perspectives. Inf. Biom. Epidemiol. in Med. u. Biol. **30**, 87–95.

ENDRENYI, L., AMIDON, G. L., MIDHA, K. K., SKELLY, J. P. (1998): Individual bioequivalence: attractive in principle, difficult in practice. Pharm. Res. **15**, 1321–1325.

ENDRENYI, L., TOTHFALUSI, L. (1999): Subject-by-formulation interaction in determinations of individual bioequivalence. Bias and Prevalence. Pharm. Res. **16**, 186–190.

ESINHART, J. D., CHINCHILLI, V. M. (1994): Extension to the use of tolerance intervals for the assessment of individual bioequivalence. J. Biopharm. Stat. **4**, 39–52.

FOOD AND DRUG ADMINISTRATION, Center for Drug Evaluation and Research, 1997: Guidance for industry: In vivo bioequivalence studies based on population and individual bioequivalence approaches. Federal Register **62**, no. 249, Dec 30, 1997.

FOOD AND DRUG ADMINISTRATION, Center for Drug Evaluation and Research, 1998a: Bioequivalence studies. http://www.fda.gov/cder/bioequivdata/index.htm.

FOOD AND DRUG ADMINISTRATION, Center for Drug Evaluation and Research, 1998b: Statistical methods for obtaining confidence intervals for individual and population bioequivalence criteria. http://www.fda.gov/cder/bioequivdata/statproc.htm.

GRAHNÉN, A., HAMMERLUND, M., LUNDQVIST, T. (1984): Implications of intraindividual variability in bioavailability studies of furosemide. Eur. J. Clin. Pharmacol. **27**, 595–602.

LIU, J. P., CHOW, S. C. (1997): A two one-sided tests procedure for assesment of individual bioequivalence. J. Biopharm. Stat. **7**, 49–61.

PIGEOT, I. (1999): Comments on bootstrap confidence intervals for evaluating bioequivalence criteria. Inf. Biom. Epidemiol. in Med. u. Biol. **30**, 96–109.

SHUMAKER, R. C., METZLER, C. M. (1998): The phenytoin trial is a case study of 'individual' bioequivalence. Drug Inform. J. **32**, 1063–1072.

Address for correspondence: Dr. Günther Pabst, Biometrie, AAI Deutschland GmbH & Co KG, Wegenerstraße 13, D-89231 Neu-Ulm, Germany. Tel.: +49-731-9840-340, Fax: +49-731-9840-358, e-mail: Guenther.Pabst@aai.de

Informatik, Biometrie und Epidemiologie in Medizin und Biologie **30** (3),
122–130, ISSN 0943-5581

Some Comments on a Recent FDA Draft Guidance on Bioequivalence Assessment

Kommentare zur Draft-Version einer neuen FDA-Richtlinie zum Nachweis der Bioäquivalenz

J. Röhmel

Summary

This paper makes some comments on a draft FDA guidance on bioequivalence assessment which was recently released for comments. The proposed metrics, bioequivalence criteria (for individual and population bioequivalence), and design are discussed. The statistical model is derived, and its relation to the statistical assessment of average bioequivalence from repeated cross-over is outlined.

Keywords

FDA guidance, average bioequivalence, population bioequivalence, individual bioequivalence, statistical model

Zusammenfassung

Es wird die Draft-Version der FDA-Richtlinie zum Nachweis der Bioäquivalenz kommentiert. Insbesondere wird auf die vorgeschlagenen Metriken, auf Bioäquivalenzkriterien (für „individuelle" und „Populations"-Äquivalenz) und auf das vorgeschlagene Design eingegangen. Das statistische Modell für solche Studien wird entwickelt, und die Beziehungen zur „mittleren" Bioäquivalenz, wie sie aus Crossover-Studien mit mehrfachen Messungen ermittelt werden kann, werden hergestellt.

Stichworte

FDA-Richtlinie, „mittlere" Bioäquivalenz, „Populations"-Bioäquivalenz, „individuelle" Bioäquivalenz, statistisches Modell

1 Introduction

In October 1997, the FDA has issued a draft guideline on the assessment of bioequivalence. This draft guideline is expressing the intention to replace the current standard procedure of demonstrating a*verage bioequivalence* by more sophisticated procedures such as *population bioequivalence* prior to approval of an originator formulation (prescribability), or *individual bioequivalence* for generic formulations or for changes after approval (switchability).
Although there is a considerable agreement among biostatisticians that *average bioequivalence* is a rather weak condition to be satisfied by a generic formulation prior to approval, there is not much evidence that the practice used so far has been inadequate, and at least in Europe there are only few (if any) real examples showing this inadequacy. It is the purpose of this paper to discuss and to comment on important proposals made in this draft guideline, e.g. metrics, bioequivalence criteria, consequences of the new proposals on the study design and on the statistical analysis. Furthermore, I will provide some technical details of the relationship between the new concepts of individual and population bioequivalence and the average bioequivalence approach using a repeated cross-over design.

2 Metrics

For single dose studies, the proposed five metrics ($AUC_{0-\tau}$, $AUC_{0-\infty}$, elimination rate constant λ_Z, elimination half-life $t_{1/2}$, peak concentration C_{max}, and time to peak concentration t_{max}) are the usual ones. Among these 5 variables, $AUC_{0-\infty}$ and C_{max} have traditionally carried a larger weight with regard to the decision of bioequivalence.
For multiple dose studies, six criteria are mentioned. All of them have been in use in this kind of study: $AUC_{0-\tau}$, C_{max}, T_{max}, C_{min}, average drug concentration at steady state C_{av}, and degree of fluctuation DF.

3 Bioequivalence criteria

There have different criteria been discussed in the literature. FDA is now proposing 'scaled' moment based criteria (e.g. in contrast to unscaled or probability based criteria). For population equivalence the following criterion is suggested:

$$\frac{(\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2)}{\max\{0.04, \sigma_{TR}^2\}} \leq \theta_P .$$

where

$$\theta_P = \frac{(\ln 1.25)^2 + 0.02}{0.04} = 1.74 .$$

Here the μ_T, μ_R, represent the true (unknown) population averages of the test and the reference formulation, σ_{TT}^2 is the total (within subject + between subject) variance of the test formulation, and σ_{TR}^2 has the respective meaning for the reference formulation. Furthermore, $\max\{0.04, \sigma_{WT}^2\}$ abbreviates the maximum of $\sigma_{W0}^2 = 0.04$ and σ_{WT}^2, where $\sigma_{W0} = 0.2$ is a constant suggested by the FDA.
Comment 1: The determination of the used constants 0.02, 0.04, and θ_P (=1.74) is derived in Appendix A of the FDA draft guidance. Their derivation is based on few studies in a FDA data base. One should consider these constants as preliminary until more experience with data of this type has been gained, particularly in Europe.

For individual bioequivalence, a similar criterion has been suggested:

$$\frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + (\sigma_{WT}^2 - \sigma_{WR}^2)}{\max\{0.04, \sigma_{WT}^2\}} \leq \theta_I ,$$

where θ_I should be a value between 2.24 and 2.49.
Here, σ_{WT}^2 is the within subject variance of the test formulation, σ_{WR}^2 has the respective meaning for the reference formulation, and σ_D^2 is the subject-by-formulation interaction variance component. Appendix A of the guidance contains some arguments supporting the suggested determination of the equivalence limit θ_I

$$\left(\frac{(\ln 1.25)^2 + 0.04}{0.04} = 2.24 \leq \theta_I \leq \frac{(\ln 1.25)^2 + 0.05}{0.04} = 2.49 \right).$$

Comment 2: As outlined above, the used constants should be considered as values one can work with until more experience with data of this type has been gained. Both criteria, for population and for individual bioequivalence award the test drug, if its variability is smaller than that of the reference drug. Smaller variability of the test drug can even offset some existing difference between test drug and reference in the population averages, but is this really desirable?

4 Study design

A) Experimental design
For *population bioequivalence*, the conventional two-formulation, two-period cross-over may still be used. Also, other forms of design (e.g. parallel groups or replicated cross-over) are still possible.
Comment 3: The requirement to demonstrate population bioequivalence would not impose major changes in the current practice of showing bioequivalence, and in particular, the traditional design of a two-sequence, two-period, two-formulation cross-over (see, e.g. JONES and KENWARD, 1989) can be used. The requirement for population equivalence in addition to average bioequivalence would mean similarity (equivalence) of the population distributions (assuming log-normal distribution of the respective pharmacokinetic parameter), because a normal distribution is completely characterized by its mean and its variance. There exist of course criteria to verify this through alternative methods compared to the proposed single aggregate criterion (incorporating both, differences in means and differences in variability), such as two separate, disaggregate criteria (one criterion for the difference in means and another for evaluating differences in variability). Again experience in the regulatory setting with aggregate or disaggregate criteria is almost completely lacking. Quite obviously, no aggregate criterion can guarantee both, average bioequivalence and equivalence of variances. Therefore, it may in some (yet unknown) situations happen, that average bioinequivalence **and** population and/or individual bioequivalence may be true at the same time.
For *individual bioequivalence* a design has to be used which allows to estimate the within formulation variance. Therefore each formulation has to be repeated for at least some study participants. The most straightforward designs are (T = Test, R = Reference):

- the two-formulation, three-period, two-sequence design, e.g. (TRT, RTR), or (TTR, RRT), or ...
- the two-formulation, four-period, two-sequence design, e.g. (TRTR, RTRT), or (TTRR, RRTT), or ...

Comment 4: Designs with more than two sequences (e.g. such as TRTR, RTRT, TRRT) are not recommended by the FDA except for special situations. This general warning is

not supported by some references, but seems to rely on unpublished work of the FDA task force group.

B) Sample size
Comment 5: Currently, sample size formulas for the proposed criteria do not exist. Hence, simulation studies are necessary. If prior estimates (guesses) for the needed variance terms (e.g. σ^2_{WT}, σ^2_{BT}, σ^2_D) can be obtained from the published literature this will make simulation straightforward. Otherwise estimates must be collected from pilot experiments. It is however important to mention that the estimates from pilot experiments should not be taken at face value.

C) Dropouts
Usually and independently from the type of study, the dropout issue is a difficult one. Therefore, it is not surprising that statements regarding dropouts in the FDA draft guidance are not particularly clear. For example, on the one hand should dropouts not be replaced during the course of the study, on the other hand should any intention to recruit additional subjects as a replacement for dropouts indicated in the protocol. Also the use of available data from dropouts is not clear. Data from dropouts may complicate the analysis only if the (incomplete) data are included. Usually participants drop out before a certain period starts, but having data on all previous periods. Only occasionally it may happen that participants miss an intermediate period. Data from dropouts with more than only the first period may be included (assuming an experimental design with alternating treatments (TRTR, RTRT)). Data from dropouts with the first period only should be excluded from the analysis. A statement should be included that no analysis is complete unless the potential biases arising from these specific exclusions, or any others, are addressed (from ICH E9 Statistical Principles for Clinical Trials; ICH, 1998).
Data of replacement subjects, if assayed, should be used for the analysis, because it increases both the chance to find a truly existing bioequivalence or a truly existing bioinequivalence.

5 Statistical analysis

A) Logarithmic transformation
The advantage of logarithmic transformation lies in the fact that readily available statistical software can be used to analyze the data. However, the proposal in favor of logarithmic transformations seems to be too strong. Its mentioning as a currently acceptable procedure would be enough. There is no principal obstacle in statistical methodology to use other procedures on the original data which yield estimates and confidence intervals for the ratio of two means and for other criteria. The development of such procedures is necessary. This necessity arises, for example as a consequence of the proposal that t_{max} and DF (degree of fluctuation) should not be transformed logarithmically. What is important is that the selected procedure is laid down in the study protocol.

B) Statistical model
For an experimental design with four-periods, two-sequences, and two-formulations T and R, the following model is proposed:
Subjects should be randomly assigned to the two sequences: T R T R (sequence 1), and R T R T (sequence 2). Subjects can then be identified by a subscript with two components, indicating the sequence and the number within sequence: $(1, 1), \ldots, (1, n_1)$ in sequence 1; $(2, 1), \ldots, (2, n_2)$ in sequence 2. Usually n_1 will equal n_2 at the planning stage, but differences in these numbers may occur if there are dropouts. The value of the

particular pharmacokinetic response variable (possibly logarithmically transformed) is denoted by X_{ijv}, which represents the value of this variable in subject j $(1 \leq j \leq n_i)$ of sequence i $(i = 1, 2)$, and period v $(v = 1, 2, 3, 4)$.
It may help understanding to write down the statistical model separately for subjects in sequence 1 and in sequence 2:

sequence 1:	sequence 2:
$x_{1j1} = \mu_{T1j} + S_1 + \pi_1 + \in_{1j1}$	$x_{2k1} = \mu_{R2k} + S_2 + \pi_1 + \in_{2k1}$
$x_{1j2} = \mu_{R1j} + S_1 + \pi_2 + \in_{1j2}$	$x_{2k2} = \mu_{T2k} + S_2 + \pi_2 + \in_{2k2}$
$x_{1j3} = \mu_{T1j} + S_1 + \pi_3 + \in_{1j3}$	$x_{2k3} = \mu_{R2k} + S_2 + \pi_3 + \in_{2k3}$
$x_{1j4} = \mu_{R1j} + S_1 + \pi_4 + \in_{1j4}$	$x_{2k4} = \mu_{T2k} + S_2 + \pi_4 + \in_{2k4}$
$j = 1, \ldots, n_1$	$k = 1, \ldots, n_2$

There are additional constrains and assumptions on fixed and random effects:

(i) fixed effects:

- $S_1 + S_2 = 0$
- $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 0$

(ii) random effects:

- $(\mu_{T1j}, \mu_{R1j}), (\mu_{T2k}, \mu_{R2k}) \underset{iid}{\approx} N((\mu_T, \mu_R), \Sigma)$ with $\Sigma = \begin{bmatrix} \sigma^2_{BT} & \varrho\sigma_{BT}\sigma_{BR} \\ \varrho\sigma_{BT}\sigma_{BR} & \sigma^2_{BR} \end{bmatrix}$
- $(\in_{1j1}, \in_{1j3}, \in_{1j2}, \in_{1j4}), (\in_{2k2}, \in_{2k4}, \in_{2k1}, \in_{2k3}) \underset{iid}{\approx} N((0, 0, 0, 0), \Omega)$

$$\Omega = \begin{bmatrix} \sigma^2_{WT} Id & 0_{2\times 2} \\ 0_{2\times 2} & \sigma^2_{WR} Id \end{bmatrix}, \qquad Id = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad 0_{2\times 2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Further details about the statistical model, the estimation of parameters, and the relationship to the assessment of average bioequivalence from repeated cross-over designs are given in the appendix.
Comment 6: This is the model which has been used frequently in the literature to analyze data from a replicative cross-over. So far, the main purpose when using such a design has been to increase precision of the estimates because of the replicative measurements for each formulation. Therefore, the use of such a design to assess average bioequivalence is known to be connected with a decrease of the number of subjects needed (CHINCHILLY, ESINHART, BARR, 1997). As a means to assess population bioequivalence or individual bioequivalence it has rarely been used, and at least in Europe, experience is thin.

C) Data analysis
Comment 7: It should be avoided to mention a particular commercial software in a paper issued by a neutral regulatory office. Other commercial software products may have similar procedures which may perform well. It is not particular to the mentioned software to be able to perform the required calculations, and obviously, additional individually written macros are necessary to extract the desired information.
Comment 8: A crucial point is the recommendation to derive bounds for the confidence intervals of the proposed bioequivalence criteria via bootstrapping. Different from the situation when analytic formulas exist for the calculation of confidence bounds it will almost surely never happen that two different experimenter will end at identical confi-

dence bounds, because bootstrapping contains an element of chance. In principle, the situation is similar to the that of randomized tests in which the result of throwing a coin will determine the p-value. This kind of tests have never been accepted in the regulatory setting. Therefore, one should require that the number of bootstraps is large enough to achieve stability within *a priori* defined margins. Whether the proposed 2000 bootstraps can regularly achieve this has to be investigated. I doubt.
Also, some additional details on how to perform bootstrapping appropriately or at least some references with specific hints should be provided.

6 Miscellaneous issues

A) Carry-over effects
The terms and conditions for which the occurrence of carry-over effects is considered unlikely is acceptable. Nevertheless, a tutorial or seminar on the analysis of replicated cross-over in the presence of some kind of cross-over would be of value.

B) Outlier consideration
Considerations about outlying values are not part of the present CPMP Note for Guidance 'Investigation of Bioavailability and Bioequivalence' (CPMP, 1991). The respective part of the discussion paper represents a useful addition, and the inclusion of such a discussion should be considered for inclusion in a revised European Guideline (CPMP, 1998).
Consistent with the definition of outlying values in the FDA draft guidance is a subject who has unusual high response in all periods. This would indicate that the subject is a member of a subpopulation of possibly slow metabolisers. The effect of such a subject on the analysis could be an increase in both intra- and intersubject variability.
The omission of subjects with outlying values from the analysis is critical, and a justification for doing so should be based on strong scientific evidence. Potential outlying values as a consequence of protocol violations could, for example, be identified before assaying the samples.

Appendix

Sequence 1: T R T R	Sequence 2: R T R T
Subjects	
$(1, 1), \ldots, (1, n_1)$ in sequence 1;	$(2, 1), \ldots, (2, n_2)$ in sequence 2
$x_{1j1} = \mu_{T1j} + S_1 + \pi_1 + \in_{1j1}$	$x_{2k1} = \mu_{R2k} + S_2 + \pi_1 + \in_{2k1}$
$x_{1j2} = \mu_{R1j} + S_1 + \pi_2 + \in_{1j2}$	$x_{2k2} = \mu_{T2k} + S_2 + \pi_2 + \in_{2k2}$
$x_{1j3} = \mu_{T1j} + S_1 + \pi_3 + \in_{1j3}$	$x_{2k3} = \mu_{R2k} + S_2 + \pi_3 + \in_{2k3}$
$x_{1j4} = \mu_{R1j} + S_1 + \pi_4 + \in_{1j4}$	$x_{2k4} = \mu_{T2k} + S_2 + \pi_4 + \in_{2k4}$
$j = 1, \ldots, n_1$	$k = 1, \ldots, n_2$

Constrains:

- $S_1 + S_2 = 0$ or $S = S_1 = -S_2$.
- $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 0$

Assumptions about distributions:

- $(\mu_{T1j}, \mu_{R1j}), (\mu_{T2k}, \mu_{R2k}) \underset{iid}{\approx} N((\mu_T, \mu_R), \Sigma), \Sigma = \begin{bmatrix} \sigma_{BT}^2 & \varrho\sigma_{BT}\sigma_{BR} \\ \varrho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 \end{bmatrix}$
- $(\in_{1j1}, \in_{1j3}, \in_{1j2}, \in_{1j4}), (\in_{2k2}, \in_{2k4}, \in_{2k1}, \in_{2k3}) \underset{iid}{\approx} N((0, 0, 0, 0), \Omega)$

$$\Omega = \begin{bmatrix} \sigma_{WT}^2 Id & 0_{2\times 2} \\ 0_{2\times 2} & \sigma_{WR}^2 Id \end{bmatrix}, \qquad Id = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad 0_{2\times 2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Transformations:

$$\bar{x}_{T1j} = 0.5 \cdot (x_{1j1} + x_{1j3}) = \mu_{T1j} + S_1 + 0.5 \cdot (\pi_1 + \pi_3) + 0.5 \cdot (\in_{1j1} + \in_{1j3})$$
$$\tilde{x}_{T1j} = (x_{1j1} - x_{1j3}) = (\pi_1 - \pi_3) + (\in_{1j1} - \in_{1j3})$$

etc.

$$\begin{aligned} \bar{x}_{T1j} &= \mu_{T1j} + S + \bar{\pi}_0 + \bar{\in}_{T1j} & \bar{x}_{T2k} &= \mu_{T2k} - S - \bar{\pi}_0 0 \bar{\in}_{T2k} \\ \tilde{x}_{T1j} &= \bar{\pi}_0 + \tilde{\in}_{T1j} & \tilde{x}_{T2k} &= \tilde{\pi}_e + \tilde{\in}_{T2k} \\ \bar{x}_{R1j} &= \mu_{R1j} + S - \bar{\pi}_0 + \bar{\in}_{R1j} & \bar{x}_{R2k} &= \mu_{R2k} - S + \bar{\pi}_0 + \bar{\in}_{R2k} \\ \tilde{x}_{R1j} &= \tilde{\pi}_e + \tilde{\in}_{R1j} & \tilde{x}_{R1j} &= \bar{\pi}_0 + \tilde{\in}_{R1j} \end{aligned}$$

$$\bar{\pi}_0 = 0.5 \cdot (\pi_1 + \pi_3), \quad \text{and contrasts} \quad \tilde{\pi}_o = \frac{1}{\sqrt{2}} (\pi_1 - \pi_3), \quad \text{and} \quad \tilde{\pi}_e = \frac{1}{\sqrt{2}} (\pi_2 - \pi_4)$$

(definitions of the error terms analogously)

$$\begin{pmatrix} \bar{x}_{T1j} \\ \bar{x}_{R1j} \\ \tilde{x}_{T1j} \\ \tilde{x}_{R1j} \end{pmatrix} \approx N\left(\begin{pmatrix} \mu_T + S + \bar{\pi}_0 \\ \mu_R + S - \bar{\pi}_0 \\ \tilde{\pi}_0 \\ \tilde{\pi}_e \end{pmatrix}, \begin{bmatrix} \sigma_{BT}^2 + 0.5\sigma_{WT}^2 & \varrho\sigma_{BT}\sigma_{BR} & 0 & 0 \\ \varrho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 + 0.5\sigma_{WR}^2 & 0 & 0 \\ 0 & 0 & \sigma_{WT}^2 & 0 \\ 0 & 0 & 0 & \sigma_{WR}^2 \end{bmatrix} \right),$$

$$\begin{pmatrix} \bar{x}_{T2k} \\ \bar{x}_{R2k} \\ \tilde{x}_{T2k} \\ \tilde{x}_{R2k} \end{pmatrix} \approx N\left(\begin{pmatrix} \mu_T - S - \bar{\pi}_0 \\ \mu_R - S + \bar{\pi}_0 \\ \tilde{\pi}_e \\ \tilde{\pi}_0 \end{pmatrix}, \begin{bmatrix} \sigma_{BT}^2 + 0.5\sigma_{WT}^2 & \varrho\sigma_{BT}\sigma_{BR} & 0 & 0 \\ \varrho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 + 0.5\sigma_{WR}^2 & 0 & 0 \\ 0 & 0 & \sigma_{WT}^2 & 0 \\ 0 & 0 & 0 & \sigma_{WR}^2 \end{bmatrix} \right),$$

$$\Sigma^* = \begin{bmatrix} \sigma_{BT}^2 + 0.5\sigma_{WT}^2 & \varrho\sigma_{BT}\sigma_{BR} \\ \varrho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 + 0.5\sigma_{WT}^2 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

Estimators: Z, $\tilde{X}$, $\tilde{S}$

$$Z = \begin{pmatrix} Z_T \\ Z_R \\ Z_S \\ Z_\pi \end{pmatrix} = \begin{pmatrix} .5(\bar{x}_{T1\bullet} + \bar{x}_{T2\bullet}) \\ .5(\bar{x}_{R1\bullet} + \bar{x}_{R2\bullet}) \\ .25(\bar{x}_{T1\bullet} + \bar{x}_{R1\bullet} - (\bar{x}_{T2\bullet} + \bar{x}_{R2\bullet})) \\ .25(\bar{x}_{T1\bullet} - \bar{x}_{R1\bullet} - (\bar{x}_{T2\bullet} - \bar{x}_{R2\bullet})) \end{pmatrix}, \quad \bar{x}_{T1\bullet} = \sum_{j=1}^{n_1} \bar{x}_{T1j}, \quad \text{etc.}$$

$$Z \approx N\left(\begin{pmatrix}\mu_T\\ \mu_R\\ S\\ \bar{\pi}_0\end{pmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12} & \Sigma_{22}\end{bmatrix}\right),$$

$$\Sigma_{11} = 0.25\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\Sigma^*, \qquad \Sigma_{12} = 0.125\left(\frac{1}{n_1}-\frac{1}{n_2}\right)\Sigma^*\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix},$$

$$\Sigma_{22} = 0.0625\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\Sigma^*\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix},$$

$$\begin{pmatrix}\bar{x}_{T1\bullet}\\ \bar{x}_{R1\bullet}\\ \bar{x}_{T2\bullet}\\ \bar{x}_{R2\bullet}\end{pmatrix} \approx N\left(\begin{pmatrix}\bar{\pi}_0\\ \bar{\pi}_e\\ \bar{\pi}_e\\ \bar{\pi}_0\end{pmatrix}, \begin{bmatrix}\frac{1}{n_1}\sigma^2_{WT} & 0 & 0 & 0\\ & \frac{1}{n_1}\sigma^2_{WR} & 0 & 0\\ 0 & 0 & \frac{1}{n_2}\sigma^2_{WT} & 0\\ 0 & 0 & 0 & \frac{1}{n_2}\sigma^2_{WR}\end{bmatrix}\right),$$

$$\tilde{S} = \begin{bmatrix}S_{TT} & S_{TR}\\ S_{TR} & S_{RR}\end{bmatrix} \approx \text{Wishart}\,(n_1+n_2-2, \Sigma^*),$$

$$S_{TT} = \sum_{i=1}^{2}\sum_{r=1}^{n_i}\left(\bar{x}_{Tir}-\bar{x}_{Ti\bullet}\right)^2, \qquad S_{TR} = \sum_{i=1}^{2}\sum_{r=1}^{n_i}\left(\bar{x}_{Tir}-\bar{x}_{Ti\bullet}\right)\left(\bar{x}_{Rir}-\bar{x}_{Ri\bullet}\right),$$

$$S_{RR} = \sum_{i=1}^{2}\sum_{r=1}^{n_i}\left(\bar{x}_{Rir}-\bar{x}_{Ri\bullet}\right)^2,$$

$$S_T^2 = \sum_{i=1}^{2}\sum_{r=1}^{n_i}\left(\tilde{x}_{Tir}-\tilde{x}_{Ti\bullet}\right)^2 \approx \sigma_T^2\chi^2_{n_1+n_2-2},$$

$$S_R^2 = \sum_{i=1}^{2}\sum_{r=1}^{n_i}\left(\tilde{x}_{Rir}-\tilde{x}_{Ri\bullet}\right)^2 \approx \sigma_R^2\chi^2_{n_1+n_2-2}.$$

Usual 'average' bioequivalence

$$Z = Z_T - Z_R;\ \hat{S}^2 = 0.25\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{1}{n_1+n_2-2}\left(S_{TT}+S_{RR}-2S_{TR}\right).$$

Finally, $Z/\hat{S}^2$ follows a t-distribution with n_1+n_2-2 degrees of freedom.

Acknowledgement

I would like to express my gratitude to Dr. MEINHARD KIESER, Karlsruhe, for his substantial help in getting this manuscript to its final form, and to the anonymous reviewers for their useful suggestions to improve the paper.

References

CHINCHILLY, V. M., ESINHART, J. D. BARR, W. H. (1997): Analysis of multiple-dose bioequivalence studies. J. Biopharm. Statist. **4**, 423–435.

CPMP (1991): Investigation of Bioavailability and Bioequivalence (Note for Guidance III/54/89-EN).

CPMP (1998): Note for Guidance on the Investigation of Bioavailability and Bioequivalence (CPMP/EWP/QWP/1401/98, Draft).

FDA (1997): In Vivo Bioequivalence Studies Based on Population and Individual Bioequivalence Approaches. Guidance for Industry, Center for Drug Evaluation and Research, Food and Drug Administration (Draft).

ICH (1998): Statistical Principles for Clinical Trials, ICH-Topic E9 (Step 4 Consensus Guideline, CPMP/ICH/363/96).

JONES B. J., KENWARD, M. G. (1989): Design and Analysis of Cross-Over Trials. Chapman and Hall, London.

Address for correspondence: Prof. Dr. Joachim Röhmel, Bundesinstitut für Arzneimittel und Medizinprodukte, Seestraße 10, D-13353 Berlin, Germany. e-mail: j.roehmel@bfarm.de

Informatik, Biometrie und Epidemiologie in Medizin und Biologie **30** (3),
131–140, ISSN 0943-5581

Gefahren der Standardmethoden für Meta-Analysen bei Vorliegen von Heterogenität

Some problems of the standard meta-analysis methods in the presence of heterogeneity

Sandra Ziegler, Norbert Victor

Institut für Medizinische Biometrie und Informatik, Abteilung Medizinische Biometrie,
Ruprecht-Karls-Universität Heidelberg

Zusammenfassung

In Meta-Analysen von randomisierten, zweiarmigen Therapiestudien sind im wesentlichen zwei Modelle gebräuchlich: das Modell mit festen Effekten und das Modell mit zufälligen Effekten. Für beide Modelle hat sich ein auf gewichteten Mittelwerten und einer Normalapproximation basierender statistischer Test (FE- bzw. RE-Test) etabliert. In dieser Arbeit werden die Gefahren aufgezeigt, die diese Tests mit sich bringen, wenn sie bei Vorliegen von Heterogenität zwischen den Studien zum Einsatz kommen.
Der Fehler 1. Art beider Tests wird für Meta-Analysen mit dichotomen Zielgrößen in der Gegenwart eines Modells mit zufälligen Effekten analytisch, teilweise asymptotisch, bestimmt. Kommt der FE-Test im Modell mit zufälligen Effekten zum Einsatz, so ist er extrem liberal und seine Niveauüberschreitung wächst mit der Gesamtpatientenzahl an, bis hin zu einem asymptotischen Fehler 1. Art von eins. Auch der RE-Test kann bei Vorliegen eines Modells mit zufälligen Effekten zu liberal werden. Seine Niveauüberschreitung wird ebenfalls um so größer, je mehr Patienten in die Meta-Analyse eingeschlossen sind.
Es wird eine Korrektur des RE-Tests vorgestellt, die asymptotisch für eine gegen unendlich laufende Patientenzahl das vorgegebene Niveau einhält. Abschließend wird dieser Test zusammen mit dem Fehler 1. Art des FE- bzw. RE-Tests am Beispiel einer Meta-Analyse zum Vergleich von niedermolekularem Heparin mit Standardheparin in der Thromboseprophylaxe illustriert.

Schlüsselwörter

Meta-Analyse, Modell mit festen Effekten, Modell mit zufälligen Effekten, Heterogenität, Fehler 1. Art, Normalapproximation

Summary

In meta-analyses of randomised trials for the comparison of two treatments, two models are commonly used: the fixed effects model or the random effects model. For each model, the standard test (FE-test and RE-test, respectively) is based on weighted means and a normal approximation. The purpose of this paper is to show the problems of these tests in the presence of heterogeneity.

Meta-analyses of randomised trials with binary outcome are considered. The type I error of both tests is calculated in the presence of a random effects model by means of theoretical, partially asymptotical investigations. In case a random effects model is valid, the FE-test is highly anticonservative. The anticonservatism is the larger, the more patients are included in the meta-analysis. For the number of patients growing to infinity, the type I error even tends to 1. Also the RE-test can be anticonservative when a random effects model is present. Its anticonservatism is also monotonically increasing in the number of patients.

A modification of the RE-test is presented which holds the nominal level asymptotically for the number of patients growing to infinity. Finally, the modified test as well as the type I error of the FE- and the RE-test are illustrate using a meta-analysis for the comparison of low molecular weight heparin and standard heparin in thrombosis prophylaxis.

Key words

meta-analysis, fixed effects model, random effects model, heterogeneity, type I error, normal approximation

1 Einleitung

Meta-Analysen haben in der Medizin zur Bewältigung der ständig wachsenden Informationsflut zunehmend an Bedeutung gewonnen. Sie sind ein Werkzeug zur Ermittlung des aktuellen Stands der Erkenntnisse. Die Resultate aller verfügbaren Studien zur betrachteten Fragestellung werden in einer Meta-Analyse systematisch zusammengefaßt. In der klinischen Therapieforschung sind Meta-Analysen inzwischen ein weitgehend akzeptiertes Hilfsmittel, um Evidenz zu bündeln, Hypothesen zu generieren und die Planungsphase neuer Studien zu unterstützen.

Das derzeitige Vorgehen in Meta-Analysen randomisierter Studien basiert im wesentlichen auf zwei Modellansätzen: dem Modell mit festen Effekten (*fixed effects model*, FE-Modell) und dem Modell mit zufälligen Effekten (*random effects model*, RE-Modell). Während im FE-Modell vorausgesetzt wird, daß der Therapieeffekt in allen in die Meta-Analyse eingeschlossenen Studien der gleiche ist, wird im RE-Modell angenommen, daß die Therapieeffekte der einzelnen Studien variieren; die eingeschlossenen Studien werden im RE-Modell als Zufallsstichprobe aus einer gedachten Population von Studien betrachtet.

Die Vertretbarkeit der den beiden Modellen zugrundeliegenden Annahmen wird in der Literatur intensiv und kontrovers diskutiert (PETITTI, 1994), die Eignung der gebräuchlichen statistischen Verfahren für diese Modelle wird hingegen kaum hinterfragt. Ausnahmen bilden eine unpublizierte Simulationsstudie (LARHOLT & GELBER, 1990), in der die beiden gebräuchlichen Auswertungsansätze evaluiert und ihre Robustheit gegen Modellverletzungen untersucht wird, sowie eine Untersuchung des Standardvorgehens für das FE-Modell bei Erfülltsein der Voraussetzungen dieses Modells (BÖCKENHOFF & HARTUNG, 1998).

Besonderes Interesse gilt derzeit der Frage, wie bei Vorliegen von Heterogenität zwischen den Studien verfahren werden soll. Zwar besteht inzwischen weitgehend Konsens darüber, daß eine sorgfältige Untersuchung der Ursachen der Heterogenität weitaus relevanter ist als ein globale Analyse, dennoch ist die Ansicht, ein Übergang auf das RE-Modell löse das Problem der Heterogenität, noch weit verbreitet.
In dieser Arbeit werden für Meta-Analysen mit dichotomen Zielgrößen die Eigenschaften der gängigen statistischen Verfahren (FE- bzw. RE-Test) basierend auf dem FE- bzw. RE-Modell bei Vorliegen von Heterogenität zwischen den Studien untersucht und die Gefahren beider Verfahren aufgezeigt; ferner wird eine Verbesserung des gebräuchlichen auf dem RE-Modell basierenden Signifikanztests vorgestellt.
Diese Arbeit ist ein Exzerpt der Dissertation der Erstautorin (ZIEGLER, 1999), auf die für weitere Details verwiesen wird.

2 Modelle und Standardmethoden

Es seien k unabhängige, randomisierte, zweiarmige Therapiestudien mit dichotomer Zielgröße in eine Meta-Analyse eingeschlossen; die Therapieeffekte $\theta_1, \ldots, \theta_k$ der k Studien seien durch die Risikodifferenz, den Logarithmus des Odds Ratios oder den Logarithmus des Relativen Risikos beschrieben.
Das FE-Modell basiert auf der Annahme, daß diese Therapieeffekte fest (nicht-zufällig) und gleich sind: $\theta_1 = \ldots = \theta_k \equiv \theta$. Zusätzlich wird vorausgesetzt, daß die Schätzer für die Therapieeffekte in den einzelnen Studien $\hat{\theta}_1, \ldots, \hat{\theta}_k$ (kurz: Einzelschätzer) unabhängig sind und einer Normalverteilung um den gemeinsamen Therapieeffekt θ folgen: $\hat{\theta}_i \sim N(\theta, w_i^{-1})$ für $i = 1, \ldots, k$. In diesem Modell berechnet sich der gängige Schätzer (WHITEHEAD & WHITEHEAD, 1991) für den gemeinsamen Therapieeffekt θ als gewichteter Mittelwert der Einzelschätzer: $\hat{\theta} = \left(\sum w_i \hat{\theta}_i\right) / \left(\sum w_i\right)$. Dabei sind die Gewichte w_i gerade die Inversen der Varianzen der Einzelschätzer; diese Varianzen werden hier Einzelvarianzen genannt. Als 95%-Konfidenzintervall hat sich $\hat{\theta} \pm 1.96/\sqrt{\sum w_i}$ etabliert. Beim zweiseitigen globalen Test auf Therapieeffekt wird üblicherweise (THOMPSON, 1993) die Nullhypothese $H_0 : \theta = 0$ zum Niveau α verworfen, falls die Teststatistik

$$T = \left(\sum w_i \hat{\theta}_i\right) / \sqrt{\sum w_i}$$

betragsmäßig größer ist als das $(1 - \alpha/2)$-Quantil der Standardnormalverteilung $(c_{1-\alpha/2})$. Dieser Test wird im folgenden als FE-Test bezeichnet. Die Modellannahme gleicher Therapieeffekte kann mittels des sog. Homogenitätstests formal geprüft werden. Dabei wird die Homogenitätshypothese $\theta_1 = \ldots = \theta_k$ zum Niveau α verworfen, falls $Q = \sum w_i (\hat{\theta}_i - \hat{\theta})^2$ größer ist das $(1 - \alpha)$-Quantil der Chiquadrat-Verteilung mit $k - 1$ Freiheitsgraden (FLEISS, 1993).
Eine Verallgemeinerung der FE-Modelle stellen die RE-Modelle dar. Im 1986 von DERSIMONIAN & LAIRD für Meta-Analysen eingeführten RE-Modell wird angenommen, daß die wahren Therapieeffekte von Studie zu Studie variieren können; sie sind jetzt unabhängige, normalverteilte Zufallsvariablen mit: $\theta_1, \ldots, \theta_k \sim N(\theta, \tau^2)$. Ferner wird vorausgesetzt, daß der Schätzer für den Therapieeffekt in der i-ten Studie $\hat{\theta}_i$ $(i = 1, \ldots, k)$ normalverteilt um den wahren Therapieeffekt θ_i in dieser Studie schwankt: $\hat{\theta}_i = \theta_i + \varepsilon_i$ mit einem von θ_i unabhängigen Schätzfehler $\varepsilon_i \sim N(0, w_i^{-1})$. Unter diesen Annahmen setzt sich die Varianz jedes Einzelschätzers aus der studienspezifischen Varianz w_i^{-1} (kurz: Einzelvarianz) und der Varianz zwischen den Studien (τ^2) zusammen: $\hat{\theta}_i \sim N(\theta, w_i^{-1} + \tau^2)$. Der gemeinsame Erwartungswert der Therapieeffekte, θ, wird durch $\hat{\theta}^* = \left(\sum w_i^* \hat{\theta}_i\right) / \left(\sum w_i^*\right)$, mit $w_i^* = (w_i^{-1} + \hat{\tau}^2)^{-1}$, geschätzt; die Gewichte setzen sich hierbei aus der Einzelvarianz w_i^{-1} und dem von DERSIMONIAN & LAIRD (1986)

vorgeschlagenen Schätzer für die Varianz zwischen den Studien

$$\hat{\tau}^2 = \max\{(Q-(k-1))/(\sum w_i - \sum w_i^2/\sum w_i), 0\},$$

der im weiteren als DL-Schätzer bezeichnet wird, zusammen. Als 95%-Konfidenzintervall ist $\hat{\theta}^* \pm 1.96/\sqrt{\sum w_i^*}$ gebräuchlich. Für einen globalen Test auf Therapieeffekt wird die Nullhypothese $H_0 : \theta = 0$ verworfen, falls die Teststatistik

$$T^* = \left(\sum w_i^* \hat{\theta}_i\right) \Big/ \sqrt{\sum w_i^*}$$

betragsmäßig größer ist als $c_{1-\alpha/2}$ (BERLIN et al., 1989). Dieser Test wird hier als RE-Test bezeichnet. Zur Durchführung der vorgestellten Verfahren stehen SAS-Macros (KUSS & KOCH, 1996) zur Verfügung.
Wie üblich (BIGGERSTAFF & TWEEDIE, 1997, HARDY & THOMPSON, 1996, LARHOLT et al., 1990) wird im folgenden vorausgesetzt, daß die Einzelvarianzen *bekannt* sind, obwohl sie in der Praxis unbekannt sind und aus den Daten geschätzt werden müssen. Simulationsuntersuchungen zeigten, daß die auf der Basis dieser Annahme gewonnenen theoretischen Resultate qualitativ mit der Realität übereinstimmen.

3 Fehler 1. Art der Standardmethoden

Fehler 1. Art des FE-Tests bei Vorliegen von Heterogenität

Der FE-Test ist für Situationen konstruiert, in denen die Annahmen des FE-Modells erfüllt sind. In der konkreten Anwendungssituation ist jedoch unbekannt, ob diese Annahmen erfüllt sind. Zwar steht der Homogenitätstest zur Verfügung, um diese Annahmen zu prüfen, jedoch hat dieser Test bekanntlich geringe Power (JONES et al., 1989). Deshalb sind Robustheitsuntersuchungen zum FE-Test gegen Abweichungen von der Modellannahme wichtig. Dazu wurde der Fehler 1. Art des FE-Tests bei Vorliegen von Heterogenität, die durch ein RE-Modell beschrieben werden kann, (kurz: Fehler 1. Art des FE-Tests *im RE-Modell*) analysiert. Um festzustellen, wie sicher der Homogenitätstest in einer solchen Situation vor der unberechtigten Anwendung des FE-Tests schützen kann, wurde außerdem die Power des Homogenitätstests untersucht.
Der Fehler 1. Art des FE-Tests im RE-Modell berechnet sich zu:

$$\alpha(\text{FE-Test} \mid \text{REM}) = 2\left\{1 - \Phi\left(c_{1-\alpha/2}\left(1+\tau^2 \sum w_i^2/\sum w_i\right)^{-1/2}\right)\right\}, \qquad (1)$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet. Aus (1) ergibt sich, daß der FE-Test im RE-Modell mit $\tau^2 > 0$ *immer* zu liberal ist. Die Niveauüberschreitung wird um so größer, je größer die Varianz zwischen den Studien ist, d. h. je stärker die Heterogenität ist. Überraschenderweise steigt die Liberalität mit der Gesamtpatientenzahl (N) in der Meta-Analyse an[1]); für $N \to \infty$ konvergiert der Fehler 1. Art sogar gegen Eins. Dies macht die unkritische Anwendung des FE-Tests besonders gefährlich, weil die Erhöhung der Patientenzahl gerade der Zweck einer Meta-Analyse ist.
Die Niveauüberschreitung nimmt unter Umständen drastische Ausmaße an: Die zufälligen Wahrscheinlichkeiten für das Eintreten des Zielereignisses unter Therapie j in der i-ten Studie (p_{ij}, $i = 1, \ldots, k$; $j = 1, 2$) seien so verteilt, daß logit $(p_{ij}) \sim N(\mu, \sigma^2)$ gilt,

[1]) Hierbei wird vorausgesetzt, daß alle Quotienten $\gamma_{ij} = n_{ij}/(n_{i1} + n_{i2})$ und $\gamma_i = (n_{i1} + n_{i2})/N$ bei wachsendem N konstant bleiben ($i = 1, \ldots, k$; $j = 1, 2$), wobei n_{ij} die Patientenzahl im Therapiearm j der i-ten Studie bezeichnet. Diese Eigenschaft wird auch bei allen folgenden asymptotischen Resultaten (für $N \to \infty$) vorausgesetzt.

Tabelle 1: Fehler 1. Art des FE- und des RE-Tests und Power des Homogenitätstests bei Vorliegen eines Modells mit zufälligen Effekten

k	n^*	**Fehler 1. Art des FE-Tests**$^-$	**Fehler 1. Art des RE-Tests**†	**Power des Homogenitätstests**§
5	50	0.14	0.08	0.26
5	100	0.23	0.10	0.46
5	200	0.34	0.11	0.69
5	500	0.51	0.12	0.90
10	50	0.14	0.07	0.40
10	100	0.23	0.08	0.69
10	200	0.34	0.08	0.91
10	500	0.51	0.08	0.99

* Anzahl der Patienten pro Therapiegruppe und Studie
$^-$ nach Formel (1) berechnet
† nach Formel (2) berechnet
§ Die Power des Homogenitätstests zum Niveau $\alpha = 0.05$ berechnet sich in der vorliegenden Situation (gleiche Einzelvarianzen: $w_1 = \ldots = w_k \equiv w$) zu: $1 - F_{k-1}(\tilde{c}_{1-\alpha}/(1 + w\tau^2))$, wobei F_{k-1} die Verteilungsfunktion der Chiquadrat-Verteilung mit $k - 1$ Freiheitsgraden und $\tilde{c}_{1-\alpha}$ das zugehörige $(1 - \alpha)$-Quantil bezeichnet.

so daß mit dem Logarithmus des Odds Ratios als Maß für den Therapieeffekt ein RE-Modell mit $\theta = 0$ und $\tau^2 = 2\sigma^2$ vorliegt. Die beiden Parameter μ und σ^2 seien so gewählt, daß die Erfolgswahrscheinlichkeiten das 95%-Referenzintervall 0.2 ± 0.1 besitzen. Einige Resultate für diese Situation sind in Tabelle 1 zusammengestellt. Der Fehler 1. Art des FE-Tests beträgt z. B., unabhängig von der Anzahl der Studien, bei jeweils 100 Patienten pro Therapiegruppe 23% und bei 500 Patienten pro Therapiegruppe sogar 51%. Angesichts dieser enormen Liberalität muß vom Einsatz des FE-Tests bei Vorliegen von Heterogenität unbedingt abgeraten werden. Leider hat der Homogenitätstest eine zu geringe Power (siehe Tabelle 1), um die Anwendung des FE-Tests in heterogenen Situationen zufriedenstellend zu verhindern.

Fehler 1. Art des RE-Tests bei Vorliegen von Heterogenität

Um den fälschlichen Einsatz des FE-Tests im RE-Modell zu verhindern, empfiehlt THOMPSON (1993), stets den FE- *und* den RE-Test parallel durchzuführen, und den RE-Test als Sensitivitätsanalyse zu betrachten. Die *generelle* Anwendung des RE-Tests ist ein zweites Vorgehen zur Vermeidung der fälschlichen Anwendung des FE-Tests im RE-Modell. Bei Gelten der Voraussetzungen des FE-Modells ist dieses Vorgehen zwar in der Regel zu konservativ, aber nie zu liberal (siehe ZIEGLER, 1999). Zur Prüfung, ob diese beiden Vorgehensweisen bei Vorliegen von Heterogenität das Niveau einhalten können, wurde auch der Fehler 1. Art des RE-Tests im RE-Modell (mit $\tau^2 > 0$) untersucht.
Obwohl die Teststatistik des RE-Tests unter der Nullhypothese *nicht* der Standardnormalverteilung folgt, wird sie üblicherweise mit dem kritischen Wert dieser Verteilung verglichen. Der tatsächliche Fehler 1. Art stimmt daher nicht mit dem vorgegebenen

[2]) Diese Bedingung ist erfüllt, falls die Patientenzahlen in allen Studien sowie beiden Gruppen gleich sind und zusätzlich die Erfolgswahrscheinlichkeiten p_{i1} $(i = 1, \ldots, k)$ der gleichen Verteilung folgen und dasselbe für p_{i2} $(i = 1, \ldots, k)$ gilt.

Niveau überein. Für den Fall gleicher Einzelvarianzen[2]), d. h. $w_1 = \ldots = w_k \equiv w$, erhält man:

$$\alpha(\text{RE-Test} \mid \text{REM}) = 2\left\{1 - \Phi\left(c_{1-\alpha/2}/\sqrt{1+w\tau^2}\right) F_{k-1}\left((k-1)/(1+w\tau^2)\right)\right.$$

$$\left. - \int_{(k-1)/(1+w\tau^2)}^{\infty} \Phi\left(c_{1-\alpha/2}\sqrt{x/(k-1)}\right) f_{k-1}(x)\,dx\right\}, \quad (2)$$

wobei F_{k-1} und f_{k-1} die Verteilungs- und die Dichtefunktion der χ^2_{k-1}-Verteilung bezeichnen. Die Betrachtung von (2) zeigt, daß der RE-Test in seinem eigenen Modell das Niveau *nicht* einhält; er kann zu konservativ, erstaunlicherweise aber auch zu liberal werden. Bei geringer Heterogenität zwischen den Studien ist der Test zu konservativ, bei großer Heterogenität zu liberal. Die Liberalität ist dabei um so größer, je größer die Heterogenität ist. Die Niveauüberschreitung wächst mit der Gesamtpatientenzahl an und kann inakzeptabel groß werden. Für das im vorangegangenen Abschnitt erläuterte Beispiel wurde (2) für verschiedene Anzahlen von Studien und Patienten berechnet. Die Resultate sind in Tabelle 1 dargestellt. Der Fehler 1. Art des RE-Tests beträgt beispielsweise bei 5 Studien mit jeweils 100 Patienten pro Gruppe 10%, bei 500 Patienten pro Gruppe 12%.
Für $N \to \infty$ konvergiert der Fehler 1. Art gegen ein asymptotisches Niveau (siehe Anhang). Im Spezialfall gleicher Einzelvarianzen reduziert sich das asymptotische Niveau (3) zu

$$\lim_{N\to\infty} \alpha(\text{RE-Test} \mid \text{REM}) = 2 \cdot (1 - F_{t_{k-1}}(c_{1-\alpha/2}))\,,$$

wobei $F_{t_{k-1}}$ die Verteilungsfunktion der t-Verteilung mit $k-1$ Freiheitsgraden bezeichnet. Der asymptotische Fehler 1. Art hängt in diesem Spezialfall ausschließlich von der Anzahl der Studien ab, und er ist um so größer, je weniger Studien in die Meta-Analyse eingeschlossen sind. Bei einem vorgegebenen Niveau von 5% beträgt das asymptotische Niveau 19% für 3 Studien, 12% für 5 Studien und 8% für 10 Studien. Bei Studien unterschiedlicher Patientenzahl ist das asymptotische Niveau zum Teil erheblich größer.
Das asymptotische Niveau liegt stets über dem nominellen Niveau und nähert sich mit wachsender Anzahl von Studien immer besser dem nominellen Niveau an. Für $k \to \infty$ konvergiert das asymptotische Niveau gegen das nominelle Niveau, jedoch ist diese Asymptotik seltener von Relevanz als die Asymptotik in der Patientenzahl.

Zusammenfassend ist zu sagen, daß bei Vorliegen von Heterogenität (RE-Modell) sowohl der FE-Test als auch der eigens für dieses Modell konstruierte RE-Test zu liberal sein können. Die Niveauüberschreitung des RE-Tests ist deutlich geringer als die des FE-Tests, jedoch meist nicht vernachlässigbar. Die Liberalität beider Tests steigt mit der Anzahl der Patienten an, wodurch ihr Einsatz im Rahmen von Meta-Analysen besonders gefährlich wird. Die von Thompson vorgeschlagene Sensitivitätsanalyse schützt demnach wirkungsvoller als der Homogenitätstest vor der eklatanten Niveauüberschreitung des FE-Tests. Zeigt die Sensitivitätsanalyse einen Unterschied auf, so ist der Übergang auf den RE-Test dringend anzuraten; allerdings kann auch der RE-Test zu liberal sein, weshalb seine *generelle* Anwendung keine sichere Abhilfe gegen die Verletzung des nominellen Testniveaus ist.
Der im Anhang angegebene asymptotische Fehler 1. Art des RE-Tests kann dazu genutzt werden, den RE-Test so zu korrigieren, daß er asymptotisch das nominelle Niveau

einhält. Dabei wird die gängige Teststatistik T^* verwendet, jedoch nicht mit dem üblichen kritischen Wert $\Phi^{-1}(1-\alpha/2)$ verglichen, sondern mit dem modifizierten kritischen Wert

$$c_{\text{krit}} := \hat{G}_N^{-1}(1-\alpha/2)\,.$$

Mit $\hat{G}_N$ sei diejenige Funktion bezeichnet, die sich aus der im Anhang definierten Funktion G ergibt, indem die unbekannten Limiten l_i, $\tilde{l}_i$ (siehe Anhang) durch die bekannten Größen

$$\hat{l}_{iN} := \frac{w_{iN}}{\sum_{j=1}^{k} w_{jN}}, \qquad \tilde{\hat{l}}_{iN} := \frac{w_{iN}}{\sum_{j=1}^{k} w_{jN} - \left(\sum_{j=1}^{k} w_{jN}^2\right) \Big/ \left(\sum_{j=1}^{k} w_{jN}\right)} \quad (i = 1, \ldots, k)$$

ersetzt werden. Simulationsuntersuchungen zeigten, daß diese Modifikation (kurz: MRE-Test) das nominelle Niveau nicht überschreitet, allerdings häufig sehr konservativ ist. Erfreulicherweise ist die Konservativität des MRE-Tests aber gerade in den Situationen besonders großer Liberalität des RE-Tests eher moderat.

4 Beispiel

Die vorgestellten Ergebnisse zu den Gefahren der gängigen Meta-Analyse-Tests seien am Beispiel einer Meta-Analyse zum Vergleich von niedermolekularem Heparin mit Standardheparin in der Thromboseprophylaxe nach größeren chirurgischen Eingriffen demonstriert. In diese Meta-Analyse wurden randomisierte, doppelblinde Studien aus dem Bereich der Allgemeinchirurgie und der orthopädischen Chirurgie eingeschlossen. Als Hauptzielgrößen wurden die Inzidenz tiefer Beinvenenthrombosen für die prophylaktische Wirksamkeit und die Inzidenz von Wundhämatomen für die Sicherheit der Therapie gewählt. Eine Diskussion der Ergebnisse hinsichtlich ihrer inhaltlichen Konsequenzen findet sich in (KOCH et al., 1997).

Tabelle 2: Sicherheitsanalyse in der Meta-Analyse zum Vergleich von niedermolekularem Heparin mit Standardheparin in der Thromboseprophylaxe

	k; N	Homogenitätstest	Typ der Analyse	P-Wert*	Fehler 1. Art~
Alle Studien	24; 12221	$P = 0.008$; Power = 0.66	FE-Test	0.062	0.28
			RE-Test	0.964	0.08
			MRE-Test	0.965	0.05
Niedrigdosis-Studien	18; 9921	$P = 0.156$; Power = 0.25	FE-Test	<0.001	0.18
			RE-Test	0.021	0.10
			MRE-Test	0.060	0.05

* *P*-Wert des FE-, RE- bzw. MRE-Tests

~ Fehler 1. Art des FE-Tests und *asymptotischer* Fehler 1. Art des RE- bzw. MRE-Tests. Unter der Annahme, daß ein RE-Modell vorliegt, wurde der Fehler 1. Art des FE-Tests gemäß (1) und der asymptotische Fehler 1. Art des RE-Tests gemäß ZIEGLER 1999 (Seite 50 + 51) berechnet. Der asymptotische Fehler 1. Art des MRE-Tests beträgt nach Konstruktion stets 0.05.

Hier wird die Sicherheitsanalyse der allgemeinchirurgischen Studien dieser Meta-Analyse (24 Studien mit 12221 Patienten) zur Illustration verwendet. Als Maß für den Therapieeffekt bezüglich der Sicherheit diente der Logarithmus des Odds Ratios. Die Ergebnisse der Homogenitäts-, FE-, RE- und MRE-Tests sind in Tabelle 2 zusammengestellt.
Die FE- und die RE-Analyse aller Studien liefern widersprüchliche Resultate. Während der FE-Test einen nicht-signifikanten Trend hin zu einer Überlegenheit von niedermolekularem Heparin anzeigt, läßt der RE-Test auf eine vergleichbare Sicherheit der beiden Heparine schließen. Vertrauen in das Ergebnis des FE-Tests ist angesichts des sehr hohen Fehlers 1. Art von 28% nicht angebracht. Der Homogenitätstest liefert trotz seiner recht geringen Power von 66% eine Warnung. Auch die Sensitivitätsanalyse weist auf Zweifel am Ergebnis des FE-Tests hin und rät zum RE-Test. Die generelle Verwendung des RE-Tests ist in diesem Beispiel vertretbar. Der RE- und der MRE-Test liefern hier vergleichbare Resultate.
In der Subgruppe der Studien mit niedrigdosiertem niedermolekularem Heparin (<3400 internationale Anti-Xa-Einheiten pro Tag) liefert sowohl der FE- als auch der RE-Test ein signifikantes Ergebnis. Beide Tests haben jedoch einen Fehler 1. Art, der weit über dem vorgegebenen Niveau liegt, so daß beide Resultate nicht vertrauenswürdig sind. Der Homogenitätstest kann durch seine extrem geringe Power keinen Hinweis auf Heterogenität und somit keine Warnung vor dem unberechtigten Einsatz des FE-Tests geben. Auch die Sensitivitätsanalyse sowie der generelle Einsatz des RE-Tests können hier nicht vor einer Verletzung des nominellen Niveaus schützen, da sowohl der FE- als auch der RE-Test zu liberal sind. Der MRE-Test kann das signifikante Ergebnis der beiden anderen Tests nicht bestätigen, zeigt aber einen Trend hin zu einer überlegenen Sicherheit von niedermolekularem Heparin.

5 Diskussion und Empfehlungen

In der Mehrzahl der Meta-Analysen zum Vergleich zweier Therapien wird eine Auswertung basierend auf einem FE-Modell präsentiert, in der letzten Zeit hat der Anteil der auf dem RE-Modell basierenden Analysen zugenommen. In der Literatur wurde eine kontroverse Diskussion über die Berechtigung der Annahmen beider Modelle geführt, systematische Untersuchungen der üblichen Testmethoden fehlen weitgehend.
Die hier vorgestellten Resultate haben gezeigt, daß die *beiden* gängigen Tests (FE- und RE-Test) in Meta-Analysen mit dichotomen Zielgrößen das nominelle Testniveau erheblich verletzen können. Vom Einsatz des FE-Tests bei Vorliegen von Heterogenität ist dringend abzuraten, da er in dieser Situation das Signifikanzniveau drastisch überschreitet und zwar um so mehr, je größer die Zahl der eingeschlossenen Patienten ist. Die häufige Empfehlung (BERLIN et al., 1989, COOK et al., 1995), bei Vorliegen statistischer Heterogenität den RE-Test und sonst den FE-Test zu verwenden, ist nicht haltbar, da die Power des Homogenitätstests zu gering ist, um die fälschliche Anwendung des FE-Tests im RE-Modell zu verhindern. Da der RE-Test im eigenen Modell das Niveau überschreiten kann, ist seine generelle Anwendung ebenfalls keine Abhilfe. Seine Liberalität ist zwar erheblich geringer als die des FE-Tests, jedoch nicht vernachlässigbar. Sie wächst mit der Gesamtpatientenzahl an und ist bei kleinen Anzahlen von Studien besonders groß. Das Vorgehen, stets beide Tests im Sinne einer Sensitivitätsanalyse durchzuführen, kann demzufolge auch nicht vor einer Niveauüberschreitung schützen. Somit garantiert keine der gängigen Vorgehensweisen die Einhaltung des nominellen Niveaus.
Eine Abhilfe bietet der modifizierter RE-Test (MRE-Test), der asymptotisch (Patientenzahl $\rightarrow \infty$) das Niveau einhält; er ist allerdings häufig sehr konservativ. Gerade in den Situationen besonders großer Liberalität des RE-Tests – dies ist der Fall, wenn (a) sehr große Studien mit ähnlicher Patientenzahl und großer Varianz zwischen den Studien

oder (b) sowohl kleine als auch sehr große Studien mit großer Varianz zwischen den Studien in die Meta-Analyse eingeschlossen sind – ist seine Konservativität aber akzeptabel.
Aus diesen Gründen empfiehlt es sich, beim generellen Einsatz des RE-Tests in den oben genannten Situationen den MRE- anstelle des RE-Tests zu verwenden bzw. in der von Thompson vorgeschlagenen Sensitivitätsanalyse in den obigen Situationen zusätzlich zum FE-Test den MRE- anstelle des RE-Tests durchzuführen.

6 Literatur

BERLIN, J. A., LAIRD, N. M., SACKS, H. S. et al. (1989): A comparison of statistical methods for combining event rates from clinical trials. Stat.Med. **8**, 141–151.

BIGGERSTAFF, B. J., TWEEDIE, R. L. (1997): Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. Stat.Med. **16**, 753–768.

BÖCKENHOFF, A., HARTUNG, J. (1998): Some corrections of the significance level in meta-analysis. Biometrical Journal **40** (8), 937–947.

COOK, C. J. , SACKETT, D. L., SPITZER, W. O. (1995): Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on meta-analysis. J Clin Epidemiol **48**, 167–171.

DERSIMONIAN, R., LAIRD N. (1986): Meta-analysis in clinical trials. Contr. Clin. Trial **7**, 177–188.

FLEISS, J. L. (1993): The statistical basis of meta-analysis. Stat. Methods. Med. Res. **2**, 121–145.

HARDY, R. J., THOMPSON, S. G. (1996): A likelihood approach to meta-analysis with random effects. Stat. Med. **15,** 619–629.

JONES, M. P., O'GORMAN, T. W., LEMKE, J. H., WOOLSON, R. F. (1989): A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. Biometrics **45**, 171–181.

KOCH, A., BOUGES, S., ZIEGLER, S., DINKEL, H., DAURES, J. P., VICTOR, N. (1997): Low molecular weight heparin and unfractionated heparin in thrombosis prophylaxis after major surgical intervention: update of previous meta-analyses. British Journal of Surgery **84,** 750–759.

KUSS, O., KOCH, A. (1996): Meta-analysis macros for SAS. SSNinCSDA **22**, 325–333.

LARHOLT, K., TSIATIS, A. A., GELBER, R. D. (1990): Variability of coverage probabilities when applying a random effects methodology for meta-analysis. (UnPub)

LARHOLT, K., GELBER, R. D. (1989): Heterogeneity in meta-analysis: A simulation study of fixed effect and random effect methods. (UnPub)

PETITTI, D. B. (1994): Meta-analysis, decision analysis, and cost effectiveness analysis: methods for quantitative synthesis in medicine. Oxford, New York, Oxford University Press.

THOMPSON, S. G. (1993): Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. Stat. Methods. Med. Res. **2**, 173–192.

WHITEHEAD, A., WHITEHEAD, J. (1991): A general parametric approach to the meta-analysis of randomized clinical trials. Statistics in Medicine, **10**, 1665–1677.

ZIEGLER, S. (1999): Meta-Analyse im Modell mit festen und zufälligen Effekten. Eingereicht als Dissertation an der Universität Heidelberg.

7 Anhang

Asymptotischer Fehler 1. Art des RE-Tests im Modell mit zufälligen Effekten

Die Einzelvarianzen aus Abschnitt 2 seien hier mit w_{iN}^{-1} $(i = 1, \ldots, k)$ bezeichnet, um auf ihre Abhängigkeit von der Gesamtpatientenzahl N hinzuweisen. Es sei ferner

$$l_i := \lim_{N\to\infty} \frac{w_{iN}}{\sum_{j=1}^{k} w_{jN}} \quad \text{und} \quad \tilde{l}_i := \lim_{N\to\infty} \frac{w_{iN}}{\sum_{j=1}^{k} w_{jN} - \left(\sum_{j=1}^{k} w_{jN}^2\right) \Big/ \left(\sum_{j=1}^{k} w_{jN}\right)}$$

für $i = 1, \ldots, k$. Dann lautet der asymptotische Fehler 1. Art des RE-Tests im Modell mit zufälligen Effekten (mit $\tau^2 > 0$):

$$\lim_{N \to \infty} \alpha(\text{RE-Test} \,|\, \text{REM}) = 2 \cdot (1 - G(c_{1-\alpha/2}))\,, \tag{3}$$

wobei

$$G(z) := \frac{1}{2\pi} \cdot \int_0^\infty \Phi\left(z\sqrt{y}\right) \int_0^\infty \frac{\cos\left(\frac{1}{2}\sum_{i=1}^{k} \arctan\left(\mu_i x\right) - \frac{1}{2} yx\right)}{\prod_{i=1}^{k} \left(1 + \mu_i^2 x^2\right)^{\frac{1}{4}}}\, dx\, dy\,, \quad z \in R$$

sei und $\mu_1, \ldots, \mu_k$ die Eigenwerte der Matrix $M = \left(M_{ij}\right)_{i,j=1,\ldots,k}$ bezeichnen mit

$$M_{ij} = \begin{cases} \tilde{l}_i(1 - l_i) & \text{falls } i = j \\ -\tilde{l}_i l_j & \text{falls } i \neq j \end{cases}.$$

Korrespondenzadresse: Sandra Ziegler, Institut für Medizinische Biometrie und Informatik, Abteilung Medizinische Biometrie, Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 305, D-69120 Heidelberg, Tel: 06221564371, Fax: 06221564195
email: sandra@imbi.uni-heidelberg.de

Informatik, Biometrie und Epidemiologie in Medizin und Biologie **30** (3),
141–154, ISSN 0943-5581

Empfehlungen für die Erstellung von Studienprotokollen (Studienplänen) für klinische Studien

H. SCHÄFER[1], J. BERGER[2], K.-E. BIEBLER[3], U. FELDMANN[4], E. GREISER[5], K.-H. JÖCKEL[6], J. MICHAELIS[7], A. NEISS[8], H. H. RASPE[9], B.-P. ROBRA[10], M. SCHUMACHER[11], H.-J. TRAMPISCH[12], N. VICTOR[13], J. WINDELER[13]

[1] Institut für Med. Biometrie und Epidemiologie der Philipps-Universität Marburg
[2] Universitäts-Krankenhaus Eppendorf, Institut für Mathematik und Datenverarbeitung in der Medizin
[3] Institut für Biometrie und Medizinische Informatik der Universität Greifswald
[4] Institut für Medizinische Biometrie, Epidemiologie und Medizinische Informatik der Universitätskliniken des Saarlandes
[5] Institut für Präventionsforschung und Sozialmedizin Bremen
[6] Institut für Medizinische Informatik, Biometrie und Epidemiologie des Universitätsklinikums Essen
[7] Institut für Medizinische Statistik und Dokumentation der Universität Mainz
[8] Institut für Medizinische Statistik und Epidemiologie der Technischen Universität München
[9] Institut für Sozialmedizin der Medizinischen Universität Lübeck
[10] Institut für Sozialmedizin der Universität Magdeburg
[11] Institut für Medizinische Biometrie und Medizinische Informatik der Universität Freiburg
[12] Abteilung für Medizinische Informatik und Biomathematik der Universität Bochum
[13] Institut für Medizinische Biometrie und Informatik der Universität Heidelberg

Zusammenfassung

Es werden Empfehlungen für die Abfassung von Studienprotokollen (Studienplänen) für klinische Studien gegeben. Diese Empfehlungen sind über den Bereich der klinischen Arzneimittelprüfung hinaus anwendbar auf klinische Studien mit unterschiedlichen Fragestellungen. Besonderes Gewicht wird auf Therapiestudien, Diagnosestudien und Prognosestudien gelegt. Die Empfehlungen betreffen den Inhalt des Studienprotokolls und können auch als Checkliste für die Abfassung von Studienprotokollen dienen. Im Anhang werden methodische Grundprinzipien therapeutischer, diagnostischer und prognostischer klinischer Studien zusammengestellt. Die biometrische und klinisch-epidemiologische Methodik findet besondere Berücksichtigung bei dieser Darstellung.

Stichwörter

Klinische Studie, klinische Forschung, Therapiestudien, Diagnosestudien, Prognosestudien, Studienprotokoll, Biometrie, klinische Epidemiologie

Summary

This paper provides recommendations concerning the contents of study protocols for clinical trials. These recommendations are not limited to trials with drugs and can be applied to trials with other objectives. Special weight is given to therapeutic, diagnostic and prognostic trials. The recommendations can be used as a checklist for the writing of study protocols. In an appendix, the basic methods of therapeutic, diagnostic and prognostic trials are summarized. Special consideration is given to biostatistics and clinical epidemiology.

Keywords

Clinical trial, clinical research, therapeutic study, diagnostic study, prognostic study, study protocol, biostatistics, clinical epidemiology

Präambel

Die vorliegenden Empfehlungen wurden im Auftrag der Konferenz der Fachvertreter für Medizinische Informatik, Biometrie und Epidemiologie von einem Ad hoc-Arbeitsausschuß unter Federführung des Erstautoren erarbeitet und von der Konferenz der Fachvertreter auf der Sitzung am 14. 05. 1999 in Halle bestätigt. Die Empfehlungen werden von der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie, von der Deutschen Region der Internationalen Biometrischen Gesellschaft und der Deutschen Gesellschaft für Sozialmedizin und Prävention unterstützt.

Einleitung

Als Grundlage für die Durchführung biomedizinischer Forschung am Menschen sollte stets ein detaillierter Studienplan (nach dem Englischen „study protocol" auch als Studienprotokoll bezeichnet) vorliegen, der die Voraussetzung, die konkrete Fragestellung (Zielsetzung der Studie), das Studiendesign, die Durchführung und die Organisation und Auswertung des Vorhabens beschreibt sowie Anhaltspunkte für die Interpretation verschiedener möglicher Studienergebnisse gibt. Der Studienplan dient den an der Durchführung der Studie beteiligten Personen als verbindliche Festlegung des Studienvorgehens in Durchführung, Auswertung und Ergebnisinterpretation und als Basis für Publikationen. Er dient Gutachtergremien (zum Beispiel Ethikkommissionen) als Unterlage zur Beurteilung der Studie.

Für die klinische Prüfung von Arzneimitteln ist das Vorliegen eines Studienplans (bei der Arzneimittelprüfung meist als Prüfplan bezeichnet) rechtsverbindlich vorgeschrieben (Arzneimittelgesetz 17. 08. 1995 §40, Abs. (1) Pkt. 6). Der Prüfplan muß dem jeweiligen Stand der wissenschaftlichen Erkenntnisse entsprechen. Diese Anforderung ist in rechtskräftigen Empfehlungen konkretisiert (Bundesminister für Jugend, Familie, Frauen und Gesundheit, 1987; CPMP/ICH/135/95, 1996), in denen Inhalt und Aufbau des Prüfplans spezifiziert sind. Klinische Arzneimittelprüfungen stellen aber nur einen Teil der patientenbezogenen klinischen Forschung (klinische Studien) dar. Außer der Prüfung der

Wirksamkeit und Sicherheit von Arzneimitteln umfaßt die patientenbezogene klinische Forschung eine Vielzahl anderer Forschungsziele aus den Bereichen der Vorbeugung, Erkennung, Prognosestellung, Behandlung und Rehabilitation von Krankheiten. Wie für Studien zur klinischen Arzneimittelprüfung sollte auch für jede andere klinische Studie, schon aus ethischen Gründen, vor Studienbeginn ein ausführlicher Studienplan erstellt werden. Empfehlungen wie zum Beispiel (Bundesminister für Jugend, Familie, Frauen und Gesundheit, 1987; CPMP/ICH/135/95, 1996), die für Studien zur klinischen Arzneimittelprüfung gelten, sind nur teilweise auf Studien mit anderen Fragestellungen (siehe oben) anwendbar.

Die vorliegenden Empfehlungen für die Abfassung von Studienplänen sind allgemeiner formuliert und nicht auf Studien aus dem Bereich der klinischen Arzneimittelprüfung oder auf Therapiestudien beschränkt. Sie sind grundsätzlich anwendbar auf alle Studien, in denen Probanden oder Patienten die Beobachtungseinheit darstellen, zielen jedoch in besonderer Weise auf Studien mit klinisch-evaluativen Fragestellungen. Hierzu gehören insbesondere Studien mit dem Ziel der Evaluation von Nutzen, Risiken oder Kosten therapeutischer, präventiver, rehabilitativer, diagnostischer oder qualitätssichernder Maßnahmen, Evaluation von Früherkennungsmaßnahmen, Konstruktion oder Evaluation von Prognoseschemata sowie Studien mit ätiologischen Fragestellungen. Zu einigen dieser Zielsetzungen liegen bereits spezifische Empfehlungen vor, auf die verwiesen wird (JESDINSKY et al., 1978; KÖBBERLING et al., 1989; WICHMANN et al., 1991). Es besteht Bedarf, derartige spezifische Empfehlungen für weitere Zielsetzungen zu erstellen. Die vorliegenden Empfehlungen umfassen alle Inhalte eines Studienplans, wobei besonderes Gewicht auf der Beschreibung der epidemiologischen und biometrischen Methodik sowie den damit zusammenhängenden Aspekten des Studienplans liegt.

Bei der Erarbeitung der vorliegenden Empfehlungen wurden frühere Vorschläge anderer Autoren berücksichtigt (BERGER et al., 1988; DUTINÉ et al., 1989; Ethikkommission Marburg, 1994, 1999; VICTOR et al., 1998).

Die Empfehlungen richten sich an Wissenschaftler, die in der klinischen Forschung tätig sind, an Gutachter und Forschungsförderer. Ihre sachgerechte Anwendung setzt biometrischen und klinisch-epidemiologischen Sachverstand voraus. Die Empfehlungen können insbesondere als Grundlage für die Abfassung und die Beurteilung von Anträgen an Ethik-Kommissionen dienen. Jedwede Belastung und Risiken sind nur dann ethisch vertretbar, wenn neben der Einwilligung (informed consent) des Patienten/Probanden die bestmöglichen Voraussetzungen dafür erfüllt sind, daß aus dem Forschungsvorhaben Erkenntnisse resultieren, die neu sind oder die Zuverlässigkeit bisheriger Erkenntnisse erhöhen und die einen praktischen medizinischen Nutzen erwarten lassen. Dies setzt voraus, daß die biometrische und die klinisch-epidemiologische Methodik der Studie dem Stand der Wissenschaft entspricht. Daher benötigen Gutachter die zur Beurteilung der biometrischen und klinisch-epidemiologischen Aspekte notwendigen Angaben bzw. Unterlagen.

Spezifische Anforderungen ergeben sich an Studienvorhaben, aus denen

a) therapeutische Empfehlungen bzw. Empfehlungen zur Verhütung von Krankheiten,
b) Empfehlungen zur Diagnostik
c) Aussagen über die Prognose von Krankheiten

im Sinne von Empfehlungen für die medizinische Praxis abgeleitet werden sollen. Solche Studien werden nachfolgend als

a) Therapie- bzw. Präventionsstudien
b) Diagnosestudien
c) Prognosestudien

bezeichnet. Die wichtigsten speziellen Anforderungen an derartige Studien werden wegen deren besonderer Bedeutung im **Anhang** zusammengestellt, ohne Anspruch auf Vollständigkeit.

Inhalt des Studienprotokolls

Falls mit der Studie neben der Hauptzielsetzung weitere Fragestellungen beantwortet werden sollen, so müssen die nachfolgend aufgeführten Punkte für jede Zielsetzung getrennt spezifiziert werden.

1 Zielsetzung der Studie (Hypothesen) und Einordnung der Studie als Pilot-Studie oder Haupt-Studie

Die Fragestellung ist präzise und ausführlich zu formulieren. Dazu ist insbesondere anzugeben, um welche Art von Fragestellung es sich handelt (Therapiestudie, Diagnosestudie, Prognosestudie, ätiologische Fragestellung, usw., siehe Einleitung), auf welche Erkrankung sich die Studie gegebenenfalls bezieht und welche Therapie(n) bzw. welche diagnostischen Verfahren bzw. welche prognostischen Variablen usw. Gegenstand der Untersuchung sind. Die diesbezüglichen Angaben sollten einleitend summarisch dargestellt und in den folgenden Teilen des Studienprotokolls gemäß nachfolgenden Abschnitten präzisiert und detailliert werden.
Zur präzisen Formulierung der Zielsetzung einer Studie gehört weiterhin die eindeutige Feststellung, ob und gegebenenfalls welche Empfehlungen für die medizinische Praxis aufgrund der Studienergebnisse ausgesprochen werden sollen, oder welche Schlußfolgerungen für die medizinische Praxis, für weitere Forschung o. a. aus den verschiedenen möglichen Studienergebnissen gezogen werden sollen, gegebenenfalls unter Berücksichtigung des bisherigen Standes der Forschung und bisher publizierter Studien. Angestrebte Empfehlungen bzw. Schlußfolgerungen müssen im Studienplan explizit formuliert werden (siehe Beispiel).

Beispiel:

Eine eindeutige Formulierung der Zielsetzung einer Studie zum Vergleich eines neuen blutdrucksenkenden Medikamentes X mit einem Standard-Medikament könnte lauten: *„Je nach Studienergebnis soll eine der folgenden Empfehlungen ausgesprochen werden: a) Der Einsatz des Arzneimittels X zur Behandlung von Patienten mit ... Bluthochdruck wird empfohlen. b) Eine weitere Anwendung des Arzneimittels X bei Patienten mit Bluthochdruck innerhalb und außerhalb klinischer Prüfungen sollte unterbleiben, da das Medikament keine klinisch relevante Wirksamkeit als Antihypertensivum besitzt. c) Bis Studien mit eindeutigem Ergebnis vorliegen, sollte das Medikament weiterhin nur im Rahmen kontrollierter klinischer Studien zur Prüfung seiner blutdrucksenkenden Wirkung angewendet werden, da die vorliegende Studie zusammen mit früheren Studien keine eindeutige Aussage bzgl. der blutdrucksenkenden Wirksamkeit erlaubt"*. In ähnlicher Weise können die Schlußfolgerungen formuliert werden, die aus einer Diagnose- bzw. Prognosestudie gezogen werden sollen.

Je konkreter die beabsichtigten Schlußfolgerungen und Konsequenzen im Studienprotokoll vorformuliert werden, umso zuverlässiger kann die biometrische und methodologische Planung bzw. deren Beurteilung erfolgen. Die Wahl der Methodik und der biostatistischen Verfahren hängt essentiell von den beabsichtigten Schlußfolgerungen und Konsequenzen ab. Natürlich können die tatsächlich gezogenen Schlußfolgerungen in der Publikation von den im Studienprotokoll getroffenen Festlegungen mit Begründung abweichen, insbesondere wenn seit der Studienplanung weitere Ergebnisse anderer Studien vorliegen.
Falls es sich bei der geplanten Studie um eine Pilot-Studie handelt, ist dies im Studienprotokoll ausdrücklich festzustellen. Eine Pilot-Studie dient der Informationsgewinnung für die Planung der weiteren Forschung. Sie dient häufig zur Planung einer bestimmten klinischen Studie, die dann als Haupt-Studie bezeichnet wird. Die Planung der Pilot-

Studie ist nur möglich, wenn die medizinisch-wissenschaftliche Fragestellung der Haupt-Studie bereits klar formuliert vorliegt. Eine Pilot-Studie kann aber nicht der Beantwortung der für die Haupt-Studie formulierten Fragestellung dienen. Eine Pilot-Studie kann bereits die wesentlichen Charakteristika der Haupt-Studie enthalten. Der Studienumfang (Fallzahl, Beobachtungsdauer) muß sich jedoch ausschließlich an der Zielsetzung der Pilot-Studie orientieren und darf nicht über das zur gezielten Vorbereitung der Haupt-Studie notwendige Maß hinausgehen. Falls es sich um eine Pilot-Studie handelt, ist eine Projektskizze gemäß der vorliegenden Empfehlungen für die geplante Haupt-Studie beizufügen, mindestens in Kurzform. Zusätzlich ist die Notwendigkeit der Durchführung einer Pilotstudie zu begründen. Dazu soll dargelegt werden, welche spezifischen Informationen aus der Pilot-Studie gewonnen werden sollen, inwiefern diese Informationen für die Planung der Haupt-Studie erforderlich sind und wie diese Informationen bei der Planung der Haupt-Studie berücksichtigt werden sollen. Die Möglichkeit, andere Informationsquellen zu nutzen, ist zu diskutieren, und der Vorteil der Durchführung einer Pilotstudie ist gegenüber den damit gegebenenfalls verbundenen Risiken für die Patienten oder Probanden abzuwägen.

Beispiele:

1) Vorbereitung einer Fallzahlberechnung für die Haupt-Studie: Um die für eine statistisch abgesicherte Aussage notwendige Fallzahl berechnen zu können, müssen Vorstellungen über die Streuung der Hauptzielgröße vorhanden sein. Wenn diese zum Beispiel aus publizierten Daten nicht entnommen werden können, kann zu deren Ermittlung eine Pilot-Studie durchgeführt werden. Allerdings sollte in einem solchen Fall begründet werden, warum die zu diesem Zweck verfügbaren modernen statistischen Verfahren der adaptiven Studiendesigns mit adaptiver Fallzahlschätzung nicht statt dessen angewendet werden.
2) Überprüfung der Praktikabilität der geplanten Haupt-Studie und Identifikation von praktischen Problemen einer Studiendurchführung (Pilotphase, „Feasibility"-Studie).

Eine Pilot-Studie ist keine „Haupt-Studie unter erleichterten Bedingungen". Daher können aus dem Ergebnis einer Pilot-Studie in aller Regel keine Konsequenzen für die medizinische Praxis gezogen werden. Ausnahmen sind zu begründen. In entsprechenden Veröffentlichungen muß eindeutig der Charakter der Pilot-Studie erkennbar sein. Die Projekt-Beschreibung muß klar angeben, ob das Projekt als Pilot-Studie geplant ist.

2 Darstellung des Standes der Forschung (mit Literaturzitaten für alle getroffenen Aussagen)

Relevantes Vorwissen zur Thematik soll im Überblick und in notwendigen Details dargestellt werden mit Quellenangaben. Hierzu soll nach Möglichkeit in Form einer systematischen Übersicht (CHALMERS et al., 1995) der Stand der medizinischen Forschung über die Erkrankung, über verfügbare Therapie- und Diagnosemethoden und prognostische Faktoren dargestellt werden. Dabei sind außer einschlägigen Forschungsergebnissen auch die Art der Literatursuche, -auswahl und -synthese und die biometrischen und klinisch-epidemiologischen Methoden der einbezogenen Arbeiten zu beschreiben und hinsichtlich der Zuverlässigkeit in bezug auf das berichtete Ergebnis kritisch zu würdigen (vergleiche Anhang). Bei der Beschreibung von Therapie-, Diagnose- bzw. Prognosestudien sollen insbesondere die Ergebnisse bisher durchgeführter/publizierter klinischer Studien zu den in der Studie geprüften Therapien bzw. diagnostischen Verfahren bzw. prognostischen Merkmalen bei der jeweiligen Erkrankung beschrieben werden. Falls in dem geplanten Projekt die Wirksamkeit eines Therapie-Verfahrens durch Vergleich mit einem etablierten

Standard (also nicht mit Placebo bzw. unbehandelten Kontrollen) nachgewiesen werden soll, sind zusätzlich die Ergebnisse von Studien, die die Wirksamkeit der in der Kontrollgruppe angewendeten Standardbehandlung belegen, in quantitativer Form (Schätzungen der Therapie-Effekte) anzugeben (Literaturzitate). In diesem Zusammenhang wird auf die besondere Problematik von sogenannten Äquivalenzstudien zum Zwecke eines Wirksamkeitsnachweises hingewiesen (WINDELER et al., 1995). Falls eine Placebogruppe oder Kontrollgruppe ohne spezifische Behandlung einbezogen wird, ist anzugeben, ob es eine Standardtherapie mit belegter Wirksamkeit gibt, und es ist zu begründen, daß eine Placebobehandlung ethisch und medizinisch vertretbar ist.

3 Definition der Beobachtungseinheit und Festlegung der Ein- und Ausschlußkriterien

Beobachtungseinheiten sind die Einheiten, an denen die Meßgrößen gemäß Punkt 5 erhoben werden sollen, also in der Regel Patienten oder Probanden. Festzulegen sind die teilnehmenden Institutionen und der Patienten- bzw. Probandenzugang innerhalb der Institutionen, die Kriterien für die Aufnahme eines Patienten/Probanden in die Studie sowie die Laufzeit der Patienten-/Probandenrekrutierung (vergleiche Punkt 8). Diese Angaben sollen so ausführlich sein, daß zweifelsfrei feststeht, wann und wo welche Patienten bzw. Personen um Teilnahme an der Studie gebeten werden, wie der weitere Ablauf und die Entscheidungskriterien für die definitive Aufnahme in die Studie aussehen und wann und wie die endgültige Aufnahme eines Patienten/Probanden in die Studie dokumentiert wird (zum Beispiel zentrale telefonische Anmeldung).

4 Studien-Design

Hier ist der Studientyp anzugeben (zum Beispiel prospektive Kohortenstudie, randomisierte Therapiestudie, Fall-Kontroll-Studie, Querschnittsstudie). Zu beschreiben sind im einzelnen Zahl und Art der Vergleichsgruppen, gegebenenfalls mit Unterscheidung in Prüfgruppen und Kontrollgruppen, Verfahren der Zuteilung zu den Vergleichsgruppen bzw. Kriterien und Verfahren für die Festlegung der Gruppenzugehörigkeit, bei randomisierter Zuteilung genaue Beschreibung des Randomisierungsverfahrens im zeitlichen Ablauf, auf den einzelnen Probanden bzw. Patienten bezogene Beschreibung aller Maßnahmen inkl. Diagnostik und Behandlung, getrennt nach Vergleichsgruppen, Kriterien für den vorzeitigen Abbruch von Behandlungs- und Diagnosemaßnahmen, Maskierung („Verblindung", siehe hierzu die näheren Ausführungen im Anhang Punkt A3). Alle Verfahrensweisen im Rahmen der Studie (Patientenrekrutierung, Patientenzuteilung, Patientenbehandlung, usw.) sind so detailliert zu beschreiben, daß dadurch die Abläufe eindeutig vorgegeben sind. Bei der Beschreibung des Randomisierungsverfahrens und des zeitlichen Ablaufs einer Randomisierung sind insbesondere Angaben zu den statistischen Eigenschaften des Verfahrens wie Blockbildung (jedoch nicht die Blocklänge), Balancierung/Stratifizierung nach prognostischen Variablen erforderlich, ferner Angaben zu den an der Durchführung der Randomisierung beteiligten Institutionen mit ihren Aufgaben (zum Beispiel zentrale telefonische Randomisierung durch eine Stelle außerhalb der Klinik). Es muß festgelegt werden, wann, wo und wie das Randomisierungsergebnis für jeden einzelnen Patienten dokumentiert wird, wie es bei der Auswertung der Studie berücksichtigt werden soll (zum Beispiel Auswertung „as randomised") und wie ein zuverlässiges „concealment" erreicht werden soll (zum Begriff des „concealment" siehe Anhang Abschnitt A3).

5 Definitionen aller Merkmale bzw. Meßgrößen

Alle Merkmale bzw. Meßgrößen, die an den Beobachtungseinheiten erhoben werden sollen, sind zu definieren inkl. des Beobachtungszeitraums und der Meßzeitpunkte. Die Funktionen der Meßgrößen als Zielgrößen, Störgrößen und Einflußgrößen sind festzulegen, insbesondere ist die Haupt-Zielgröße zu definieren. Alle Merkmale, Kriterien und Meßgrößen, die als Voraussetzung für die Aufnahme in die Studie oder im Verlauf der Studie an den Beobachtungseinheiten erhoben werden sollen, müssen operational definiert werden, das heißt das Meß- und Erhebungsverfahren für jede Größe bzw. Merkmal muß eindeutig beschrieben werden. Zu den wichtigsten Größen bzw. Merkmalen, insbesondere zu den Hauptzielgrößen, sollen qualifizierende Merkmale (zum Beispiel Sensitivität, Validität, Reliabilität und andere) der eingesetzten Meß- und Erhebungsinstrumente angegeben und mit Literaturangaben belegt werden, falls es sich nicht um etablierte Größen handelt.

6 Beschreibung und Diskussion der potentiellen Störeinflüsse und Maßnahmen zu deren Kontrolle in der Studiendurchführung und/oder in der Auswertung

Mögliche Störeinflüsse (Definition siehe unten) sind zu diskutieren, und es ist anzugeben, welche Methoden angewendet werden sollen, um diese auszuschalten oder zu kontrollieren. Im Anhang dieses Leitfadens sind die wichtigsten methodischen Prinzipien zur Kontrolle von Störeinflüssen bei Therapie-, Diagnose- und Prognosestudien zusammengestellt. Abweichungen von diesen Prinzipien sind zu begründen.
Unter Störeinflüssen sind alle Quellen systematischer Fehler zu verstehen. Zum Verständnis des Begriffes des systematischen Fehlers gehe man davon aus, daß in der Studie ein Zusammenhang zwischen Einflußgrößen und Zielgrößen ermittelt werden soll, und daß das ermittelte Ergebnis von dem tatsächlich bestehenden Zusammenhang abweichen kann. Diese Abweichung wird als Fehler bezeichnet. Ein Fehler hat eine Richtung und ein Ausmaß. Systematische Fehler (im Gegensatz zu zufälligen Fehlern) sind solche, die durch Vergrößerung des Stichprobenumfangs nicht reduziert werden können.

7 Beschreibung des biometrischen Auswertungsvorgehens

Diese Beschreibung muß explizit Bezug nehmen auf die mit der Studie angestrebten Empfehlungen oder Schlußfolgerungen gemäß Punkt 1 zweiter Absatz. Zu jeder der möglichen Empfehlungen bzw. Schlußfolgerungen ist anzugeben, wie das Ergebnis der biometrischen Auswertung aussehen muß, um die jeweilige Empfehlung zu stützen. Falls dazu statistische Tests und/oder Schätzverfahren angewendet werden sollen, müssen die Prüfhypothesen (Nullhypothese und Gegenhypothese) formuliert und/oder die Parameter definiert werden, die geschätzt werden sollen. Die Beschreibung muß so ausführlich sein, daß aus dem Studienprotokoll hervorgeht und nachvollzogen werden kann, wie aus den verschiedenen an den Beobachtungseinheiten erhobenen Meßgrößen die Hypothesen geprüft bzw. die zu schätzenden statistischen Parameter ermittelt werden sollen. Dazu sind insbesondere der Zeitpunkt der Auswertung und die anzuwendenden statistischen Verfahren inkl. gegebenenfalls der Irrtumswahrscheinlichkeiten bzw. Konfidenzwahrscheinlichkeiten anzugeben, und es ist festzulegen, welche Probanden/Patienten in die Auswertung einbezogen werden und wie gegebenenfalls die Gruppeneinteilung in der Auswertung erfolgt. Bei randomisierten Therapiestudien soll die Gruppeneinteilung in der Auswertung in der Regel nach dem Ergebnis der Randomisierung (das heißt Auswertung entsprechend dem sogenannten intention-to-treat-Prinzip) erfolgen, wenn Unterschiede gezeigt werden sollen. Soll ein Äquivalenznachweis geführt werden, ist in der

Regel eine Einschränkung auf die protokollgerecht behandelten Fälle („valid cases") angebracht (vergleiche (CPMP/ICH/363/96E9, 1998)). Abweichungen von diesen Auswertungsprinzipien sind zu begründen.
Die Details der statistischen Auswertung können in einem separaten statistischen Analyseplan festgelegt werden. Nötigenfalls kann auf der Basis eines „blind review" (das heißt ohne jede Information über die Zugehörigkeit der Patienten zu einzelnen Vergleichsgruppen) dieser Analyseplan den spezifischen Erfordernissen angepaßt werden. Die Grundelemente der statistischen Auswertung (siehe oben) müssen jedoch bereits im Studienprotokoll festgestellt werden.
Falls Empfehlungen im Sinne der Zielsetzung der Studie (Punkt 1 zweiter Absatz) möglicherweise auch aufgrund von Ergebnissen einer **Zwischenauswertung** ausgesprochen werden sollen, sind die vorstehenden Festlegungen für jede Zwischenauswertung zu treffen. Zusätzlich sind die Zeitpunkte der geplanten Zwischenauswertungen und die geplanten speziellen statistischen Methoden für Zwischenauswertungen (adäquate Kontrolle der globalen Irrtumswahrscheinlichkeiten bzw. Konfidenzwahrscheinlichkeiten durch sequentielle statistische Verfahren) zu beschreiben. Der Einfluß von Zwischenauswertungen auf den weiteren Verlauf der Studie ist einzuschätzen. Es ist festzulegen, welchen Personen die Ergebnisse von Zwischenauswertungen vor der abschließenden Öffnung des Randomcodes oder vor der Veröffentlichung der Studienergebnisse zur Kenntnis gegeben werden.

8 Planung des Studienumfanges und Angabe der Bedingungen für die Beendigung der Studie sowie gegebenenfalls vorgesehene Möglichkeiten zu Anpassungen/ Änderungen des Studien-Designs im Studienverlauf

Zur Planung des Studienumfanges gehören eine nach dem Stand der biometrischen Wissenschaften durchgeführte Planung der zur Beantwortung der Fragestellung notwendigen Zahl von Patienten (Probanden) bzw. Zielereignisse (bei Überlebenszeitanalysen) in den einzelnen Vergleichsgruppen und die darauf abgestimmte Planung der Beobachtungsdauer mit Festlegung der Beobachtungs- bzw. Meßzeitpunkte für jeden Einzelpatienten. Ausnahmen sind zu begründen. Die Kriterien für die Beendigung der Studie müssen vollständig und eindeutig formuliert werden und müssen Bedingungen für den Einschluß des letzten Patienten/Probanden sowie Bedingungen für die Beendigung der Nachbeobachtung enthalten. Dies kann von einer festen Zeitdauer ab Studienbeginn oder von der Anzahl eingebrachter Patienten bzw. aufgetretener Zielereignisse abhängen oder von dem Ergebnis von Zwischenauswertungen. Im letzteren Fall sind festzulegen: Zeitpunkte, Auswertungsverfahren, Entscheidungskriterien und Entscheidungsträger (siehe Punkt 7).
In bestimmten Fällen und unter Einsatz bestimmter statistischer Planungs- und Auswertungsverfahren ist es möglich, eine Studie auch dann noch zur Beantwortung einzelner Fragestellungen heranzuziehen, wenn im Studienverlauf Änderungen des Studien-Designs vorgenommen werden (zum Beispiel vorzeitiges Schließen eines Studienarms bei mehrarmigen Studien, Anpassung der vorgesehenen Patientenzahl). Solche Studien-Designs werden als adaptive Studien-Designs bezeichnet. Dies setzt voraus, daß die Art der möglichen Design-Änderungen und die Bedingungen dafür (Zeitpunkte, Auswertungsverfahren, Entscheidungskriterien, Entscheidungsträger) im Studienprotokoll festgelegt werden und speziell dafür entwickelte statistische Verfahren zur Anwendung kommen.

9 Studienorganisation und Verantwortlichkeiten

Zeitliche Strukturierung (Definitionsphase, Vorphase, Hauptphase, Auswertephase, Berichtsphase), Auflistung *aller* an der Studie beteiligten Einrichtungen und Kooperationspartner (patienteneinbringende Institutionen, Referenzzentren, Datenzentrum/Biometrie/Epidemiologie, Entscheidungsgremien, Sponsor bzw. finanzierende Institution, usw.) mit deren Funktion und Verantwortlichkeiten, jeweils verantwortliche Personen mit Angabe von deren Qualifikation und Unterschrift dieser Personen. Organisation der Studienabläufe innerhalb jeder Institution, insbesondere Patienteneinbringung, Patientenbehandlung und studienbezogenen Dokumentation innerhalb der patienteneinbringenden Institutionen, Organisation des Informationsflusses und der Zusammenarbeit zwischen den Institutionen, Entscheidungsstrukturen für die Gesamtstudie, Organisation des Datenflusses und Datenmanagements, mit Angaben zu Hard- und Software.

10 Ethik und Datenschutz

Ethische Vertretbarkeit der angewendeten Behandlungsmethoden, gegebenenfalls der Randomisierung (Gleichwertigkeit der Behandlung nach dem Stand der Wissenschaft) und der Maskierung, Art und Umfang der Probanden-/Patientenaufklärung, Zustimmung zur Weitergabe von Daten, gegebenenfalls Zustimmung zur Einsichtnahme in Patientendokumentation zum Zwecke der „source data verification", Anonymisierung der Daten, Datenschutz- und Datensicherheitskonzept, einschließlich Angaben über Aufbewahren von Originalunterlagen. Anstelle von Einzelausführungen zum Datenmanagement (Punkt 9) und zum Datensicherheitskonzept kann die Benennung einer mit diesen Aufgaben verantwortlich betrauten biometrischen Einrichtung erfolgen unter Bezug auf Standard Operating Procedures (SOPs) dieser Einrichtung.

11 Beschreibung des geplanten Qualitätsmanagements

Standardisierung aller Verfahren und Kriterien, Schulung des Personals, Prüfung der Validität und Reabilität der eingesetzten Erhebungsinstrumente und Meßverfahren (gegebenenfalls Pilot-Studie). Art, Umfang, Ort und Zeitpunkte durchzuführender Qualitätskontrollen (zum Beispiel Monitoring), geplante externe Qualitätskontrollen (Audit), vorgesehene Maßnahmen und Kriterien für deren Auslösung, Verantwortlichkeiten. Qualitätssicherungs-Maßnahmen müssen insbesondere die Sicherung der Datenqualität von der Erfassung am Patienten bis zur Verarbeitung in der Auswertung und der planmäßigen Interpretation (siehe Punkt 1 zweiter Absatz und Punkt 7) umfassen, zum Beispiel durch unabhängige Auswertung kritischer Ergebnisse durch mehrere unabhängige Wissenschaftler oder Institutionen. Qualitätsindikatoren, die zu überwachen sind, sind insbesondere die Rekrutierungsrate (Erreichen der benötigten Patienten-/Probandenzahlen), die Einhaltung der Ein- und Ausschlußkriterien, die protokollgerechte Behandlung der Patienten, die Einhaltung der Maskierung („Verblindung"), die korrekte und vollständige Dokumentation der Merkmale bzw. Meßgrößen laut Punkt 5.

12 Maßnahmen zur Gewährleistung der Probanden-/Patientensicherheit

Erfassung von unerwünschten Ereignissen und Risiken der angewendeten medizinischen Maßnahmen, Zwischenauswertungen im Hinblick auf Wirksamkeitsunterschiede inkl. der geplanten statistischen Zwischenauswertungs-Verfahren (siehe Punkt 7), vorgesehene Meldewege, Verantwortlichkeiten, vorgesehene Maßnahmen beim Einzelpatienten und

Maßnahmen für die gesamte Studie, einschl. Kriterien für deren Auslösung, gegebenenfalls Art und Umfang von Probanden-/Patientenversicherungen.

13 Diskussion der Erfolgsaussichten

Studienbezogene Struktur- und Prozeßqualität in den beteiligten Einrichtungen, insbesondere Erreichen der Fallzahl, Kalkulation der Kosten und Finanzierung der Studie, Diskussion der Eignung des Studienkonzepts und der angewandten Methodik, auch im Vergleich zu anderen möglichen Studienansätzen, im Hinblick auf die beabsichtigten Schlußfolgerungen laut Punkt 1 zweiter Absatz. Diese Diskussion sollte außer der Frage der internen Validität auch die Frage der externen Validität des Studienkonzepts umfassen, also das Problem der Übertragbarkeit der in einem selektierten Kollektiv gewonnenen Ergebnisse auf zukünftige Patienten bzw. Probanden (zum Beispiel durch Bezug auf externe Referenzdaten). Die Diskussion zur Frage der Übertragbarkeit der Studienergebnisse sollte insbesondere die Frage der Einbeziehung beider Geschlechter und unterschiedlicher Altersgruppen in die Studie einschließen. In Form einer Schwachstellenanalyse sollten erwartete Probleme und Limitationen des Studiendesigns und der Studiendurchführung a priori dargestellt werden.

14 Einhaltung publizierter Empfehlungen und Guidelines

Berücksichtigung des Standes der Wissenschaft inkl. der biometrischen und klinisch-epidemiologischen Prinzipien, bei nicht-therapeutischen Studien gegebenenfalls in sinngemäßer Anwendung. Insbesondere wird auf die Good Clinical Practice-Richtlinien (Bundesminister für Jugend, Familie, Frauen und Gesundheit, 1987; CPMP/ICH/135/95, 1996; CPMP/ICH/363/96E9, 1998), die Empfehlung zur Publikation klinischer Studien (BEGG et al., 1996), die DFG-Denkschrift (DFG, 1997, 1998) sowie gegebenenfalls zusätzliche Anforderungen von Zulassungsbehörden verwiesen. Je nach den institutionellen Rahmenbedingungen und den Finanzierungsmodalitäten einer klinischen Studie können Anpassungen notwendig sein, wie sie zum Beispiel in den MRC-Guidelines (Medical Research Council, 1998) zu finden sind. Die wichtigsten methodischen Prinzipien therapeutischer, diagnostischer und prognostischer Studien finden sich im Anhang, in Anlehnung an (SACKETT et al., 1997).

15 Publikation der Ergebnisse

Bekanntgabe und Empfänger von Zwischenergebnissen vor der Abschlußpublikation, Zeitpunkt und Inhalt von Publikationen, Autorenschaft, Zustimmungspflicht, Angaben zur Projektförderung in den Publikationen. Es sollte in jedem Fall eine Verpflichtung zur Publikation der Ergebnisse festgehalten werden, unabhängig davon, wie die Ergebnisse ausfallen.

16 Anlagen

Dem Studienprotokoll sind die Erfassungsbögen, Formular für die Patientenaufklärung und Einverständniserklärung, Versicherungsurkunden und andere studienwichtige Unterlagen beizufügen.

Anhang:

Methodische Grundprinzipien therapeutischer, diagnostischer und prognostischer klinischer Studien

Spezifische Anforderungen ergeben sich an Studienvorhaben, aus denen therapeutische Empfehlungen, Empfehlungen zur Diagnostik oder Aussagen über die Prognose von Krankheiten im Sinne von Empfehlungen für die medizinische Praxis abgeleitet werden sollen. Die besonderen methodischen Anforderungen sind in der einschlägigen Fachliteratur dargestellt, siehe zum Beispiel (SACKETT et al., 1997) (Section 3.A.1 für Diagnosestudien, Section 3.A.2 für Prognosestudien und Section 3.A.3 und 3.A.4 für Therapiestudien) sowie die Lehrbücher (POCOCK, 1983) für Therapiestudien und (KÖBBERLING et al., 1991) für Diagnosestudien und die Arbeit (SIMON et al., 1994) für Prognosestudien. Die wichtigsten methodischen Prinzipien werden nachfolgend zusammengestellt. **Auf diese Prinzipien sollte in einem Projektantrag Bezug genommen werden. Abweichungen von diesen Prinzipien sollen begründet werden.**

A1. Diagnosestudien

- Vergleich mit einem Referenz-Standard (möglichst „Gold-Standard") für die Diagnosestellung. Die Erhebung des Referenz-Standards kann gegebenenfalls eine Verlaufsbeobachtung der Patienten erforderlich machen.
- Begründung der Zuverlässigkeit des Referenz-Standards mit Literaturangaben
- Erhebung und Dokumentation des Ergebnisses des zu prüfenden diagnostischen Tests bzw. des diagnostischen Verfahrens ohne Kenntnis des Ergebnisses im Referenz-Standard, und umgekehrt
- repräsentative Einbeziehung solcher Patienten, für die aufgrund der Studienergebnisse die Anwendung des diagnostischen Tests empfohlen werden soll (insbesondere Repräsentativität hinsichtlich des Schweregrads der Erkrankung und der bereits erfolgten diagnostischen Vorabklärung im Sinne einer Patientenselektion aufgrund der Ergebnisse vorher durchgeführter diagnostischer Tests oder der Anamnese)
- Anwendung des Referenz-Standards auf alle Patienten unabhängig vom Ergebnis des zu prüfenden diagnostischen Tests
- Falls diagnostische Variablen aus einer größeren Variablenzahl ausgewählt werden oder falls diagnostische Variablen zu einer Diagnoseregel kombiniert werden oder falls Schwellenwerte für stetige diagnostische Variablen aufgrund der Studienergebnisse festgelegt werden, ist die Validierung an einem unabhängigen Kollektiv von nach den gleichen Methoden beobachteten, dokumentierten und ausgewerteten Patienten erforderlich. In begründeten Ausnahmefällen können Erkenntnisse zur Validität einer Diagnoseregel durch die Anwendung spezieller statistischer Resampling-Techniken gewonnen werden.
- Empfehlungen zur Anwendung des geprüften diagnostischen Tests sollen auf einer Schätzung der prädiktiven Werte (positiver und negativer prädiktiver Wert) beruhen, müssen also außer Sensitivität und Spezifität des Tests auch die Pre-Test-Wahrscheinlichkeit einbeziehen.
- hinreichende Genauigkeit der Schätzung aufgrund biometrischer Fallzahlplanung
- Als Alternative kommt das Studiendesign der randomisierten diagnostischen Studie in Frage: In der Prüfgruppe wird der in Frage stehende diagnostische Test durchgeführt, in der Kontrollgruppe wird er nicht durchgeführt, dann erfolgt die Nachbeobachtung aller Patienten mit vollständiger Erfassung eines klinisch relevanten Zielkriteriums, gemäß den Prinzipien der randomisierten Therapiestudie (siehe A3). Dieses Studiendesign bietet die höchste Aussagekraft.

A2. Prognosestudien

- Zusammenstellung einer oder mehrerer repräsentativer Kohorten von Patienten zu einem ähnlichen Zeitpunkt im Krankheitsverlauf (Krankheitsstadium)
- Ableitung von prognostischen Aussagen nur für Patienten in dem damit definierten Krankheitsstadium
- eindeutige Definition der Merkmale, die hinsichtlich ihrer prognostischen Bedeutung untersucht werden sollen, inkl. Meßvorschriften, Messung bzw. Erhebung zum Zeitpunkt des Studieneinschlusses
- ausreichend lange und vollständige Nachbeobachtung aller in die Kohorte eingeschlossenen Patienten
- klinisch relevanter Endpunkt (Zielereignis) mit Begründung der klinischen Relevanz (Literaturangaben)
- Erhebung und Dokumentation der Endpunkte ohne Kenntnis der Ausprägungen der in der Studie untersuchten Prognosefaktoren bzw. der Zugehörigkeit zu verschiedenen Prognosegruppen
- Adjustierung bezüglich bereits bekannter prognostischer Faktoren (Angabe von Art und Ergebnis der Literatur- und Datenbankrecherchen bezüglich bekannter prognostischer Faktoren)
- Falls aus den auf ihre prognostische Bedeutung geprüften Faktoren eine Teilmenge ausgewählt wird oder Faktoren zu einer Prognoseregel kombiniert werden oder falls Schwellenwerte für stetige Variablen aufgrund der Studienergebnisse festgelegt werden, ist eine Validierung der so konstruierten Prognoseregel an einer nach den gleichen Prinzipien beobachteten und dokumentierten unabhängigen Patientengruppe erforderlich. In Ausnahmefällen können Erkenntnisse zur Validität einer Prognoseregel durch die Anwendung spezieller statistischer Resampling-Techniken gewonnen werden, deren statistische Validität im Einzelfall abgesichert werden muß.
- ausreichende Genauigkeit der Schätzung der Überlebensraten (bzw. Ereignisraten) oder medianen Überlebenszeiten (bzw. Zeit bis zum Eintritt des Zielereignisses) in Abhängigkeit von der Ausprägung der untersuchten prognostischen Variablen, entsprechende statistische Fallplanung.

A3. Therapiestudien

- randomisierte Zuteilung der Patienten zu den Therapiegruppen. Falls keine Randomisierung vorgesehen ist, ist eine ausführliche Darstellung nachvollziehbarer Gründe für den Verzicht auf die Randomisierung erforderlich und eine ausführliche Erläuterung, warum trotzdem ein Erkenntnisgewinn erwartet wird.
- Begründung der klinischen Relevanz der Zielgröße (Endpunkt, siehe Hauptteil der vorliegenden Leitlinie, Punkt 5) im Hinblick auf die geplanten Konsequenzen der Studie laut Hauptteil, Punkt 1 zweiter Absatz
- Zum Nachweis der Wirksamkeit einer Therapie ist der Vergleich mit einer Kontrollgruppe von Patienten erforderlich, die entweder unbehandelt bleiben bzw. ein Placebo erhalten oder mit einer Therapie mit anerkannter Wirksamkeit (Zitate der Studien, in denen die Wirksamkeit nachgewiesen wurde) behandelt werden.
- Bevor dem Arzt und/oder dem Patienten das Ergebnis der Therapiezuteilung bekannt gegeben wird, müssen die Einwilligung des Patienten vorliegen und der Einschluß des Patienten in die Studie und das Ergebnis der Therapiezuteilung unabänderlich dokumentiert sein („Concealment").
- Festlegung der Patientenzahl mittels statistischer Planungsmethoden auf der Basis einer vorher festgelegten minimalen klinisch relevanten Differenz im Zielkriterium

zwischen den Therapiegruppen bei Prüfung auf Unterschied, bzw. eines Äquivalenzbereichs bei Prüfung auf Ebenbürtigkeit.

- Falls dies von der Form der angewandten Therapien her möglich ist, soll eine doppelblinde Studiendurchführung vorgesehen werden. Dies bedeutet, daß weder dem Patienten noch dem behandelnden Arzt noch gegebenenfalls dem Arzt, der die Therapieergebnisse beurteilt, die Therapie des jeweiligen Patienten bekannt ist. Bei sogenannten dreifachblinden Studien besitzt auch der auswertende Biometriker nicht die zur Identifikation der Therapien notwendige Information. Falls doppel- oder dreifachblinde Studiendurchführung nicht möglich ist, sind die Möglichkeiten einer einfachblinden Studiendurchführung (das heißt nur der Patient ist nicht über die Therapie informiert) sowie die Möglichkeit einer verblindeten Beurteilung der Therapieergebnisse durch einen unabhängigen Arzt (zum Beispiel bei Röntgenaufnahmen) zu diskutieren.
- geeignete Berücksichtigung bekannter prognostischer Faktoren bei der Studiendurchführung (Stratifizierung) und Auswertung (Adjustierung), Angabe von Art und Ergebnis durchgeführter Literatur- und Datenbankrecherchen bzgl. bekannter prognostischer Faktoren.
- vollständige Nachbeobachtung, Dokumentation und Auswertung aller Patienten unabhängig vom tatsächlichen Therapieverlauf, soweit möglich auch nach Protokollabweichungen
- biometrische Auswertung anhand einer klinisch relevanten Zielgröße, Darstellung der Unterschiede dieser Zielgrößen zwischen den beiden Gruppen mit Angabe eines Konfidenzintervalls für diesen Unterschied, Interpretation der Konfidenzgrenzen unter dem Aspekt der klinischen Relevanz der Effekte
- Präsentation einer Auswertung aller Patienten in der durch die Randomisierung festgelegten Therapiegruppe, unabhängig vom tatsächlichen Verlauf der Therapiedurchführung und gegebenenfalls Protokollverletzungen („as randomised", „intention-to-treat", „full analysis set"). Daneben können je nach Fragestellung zusätzliche Auswertungen notwendig sein, bei denen nur protokollgerecht behandelte Patienten in der jeweiligen Gruppe einbezogen werden („per protocol"). Ausnahmen von dem Prinzip der primären Auswertung „as randomised" ergeben sich bei manchen Studien im Rahmen der Arzneimittelentwicklung, zum Beispiel bei Auswertungen zur Beurteilung unerwünschter Ereignisse und bei Studien zum Äquivalenznachweis (siehe Hauptteil dieser Empfehlungen, Abschnitt 7).

In allen Studienformen ist eine adäquate statistische Auswertung, die eine Kontrolle bzw. Darstellung der Zufallsvariabilität des Ergebnisses (Konfidenzintervall) beinhaltet, sowie eine Planung der zur Erreichung einer vorgegebenen Genauigkeit (bei Therapiestudien: minimale klinisch relevante Differenz) notwendigen Fallzahl vorzusehen.
Bei klinischen Studien mit anderen als den vorstehend aufgeführten Zielsetzungen (zum Beispiel Studien zur Ätiologie, zu Risikofaktoren und andere) sind die vorstehend dargestellten Prinzipien in analoger Weise anzuwenden.

Literatur

BEGG, C. et al. (1996): Improving the Quality of Reporting of Randomized Controlled Trials. The CONSORT-Statement, JAMA **276** (8), 637–639.

BERGER, J., BERGMANN, K. E., GREISER, E., KEIL, U., LEHMACHER, W., SCHÄFER, H., SCHWARTZ, F. W., WICHMANN, H. E. (1988): Manual für die Planung und Durchführung epidemiologischer Studien auf dem Gebiet allergischer Krankheiten. Allergologie **11**, 479–492.

Bundesminister für Jugend, Familie, Frauen und Gesundheit (1987): Grundsätze für die ordnungsgemäße Durchführung der klinischen Prüfung von Arzneimitteln. 9. 12. 1987. Bundesanzeiger Nr. 243 vom 30. 12. 1987, 16167.

CHALMERS, I., ALTMAN, D. G. (eds.) (1995): Systematic Reviews. BMJ Publishing Group, London.

CPMP/ICH/135/95 (1996): Committee for Proprietary Medicinal Products of the European Commission (CPMP) and International Conference on Harmonization (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use, Note for Guidance on Good Clinical Practice: Consolidated Guideline.

CPMP/ICH/363/96E9 (1998): Note for Guidance on Statistical Principles in Clinical Trials.

DFG (1997, 1998): Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft".

DUTINÉ, G., GENTH, E., FRANKE, M., LAASER, U., WICHMANN, E., RASPE, H.-H., ZINK, A. (1989): Fachliche Hinweise für die Abfassung von klinisch-epidemiologischen Studienprotokollen und Förderungsanträgen. Unveröffentlicht.

Etikkommission Marburg (1994, 1999): Ratgeber für Projektleiter, die der Kommission für Ethik in der ärztlichen Forschung (Ethikkommission) des Fachbereichs Projekte vorlegen. Kommission für Ethik in der ärztlichen Forschung des Fachbereichs Humanmedizin der Philipps-Universität, Vorsitzender F. Heubel, ab 1998 G. Richter.

JESDINSKY, H. J. unter der Mitarbeit von FINK, H., VAN DE LOO, J., OBERHOFFER, G. (1978): Memorandum zur Planung und Durchführung kontrollierter klinischer Therapiestudien. Schriftenreihe der GMDS (1). Schattauer Verlag, Stuttgart.

KÖBBERLING, J., TRAMPISCH, H.-J., WINDELER, J. (1989): Memorandum zur Evaluierung diagnostischer Maßnahmen. Schriftenreihe der GMDS (10). Schattauer Verlag, Stuttgart.

KÖBBERLING, J., RICHTER, K., TRAMPISCH, H. J., WINDELER, J. (1991): Methodologie der medizinischen Diagnostik. Springer-Verlag, Berlin.

Medical Research Council (1998): MRC Guidelines for Good Clinical Practice in Clinical Trials. MRC Clinical Trials Series.

POCOCK, S. J. (1983): Clinical Trials. A Practical Approach. John Wiley & Sons Ltd., Chicester.

SACKETT, D. L., RICHARDSON, W. S., ROSENBERG, W., HAYNES, R. B. (1997): Evidence-based Medicine. Churchill Livingstone, New York.

SIMON, R., ALTMAN, D. G. (1994): Statistical aspects of prognostic factor studies in oncology. Br J Cancer **69**, 979–985.

VICTOR, N., HOLLE, H. (1998): Gliederungsvorschlag für Studienprotokolle bei vergleichenden Therapiestudien. In: RASCH, D. et al. (Hrsg.): Verfahrensbibliothek – Versuchsplanung und Auswertung. Band II. Oldenbourg Verlag, München 1998, pp 655–662.

WICHMANN, H. E., LEHMACHER, W., unter der Mitarbeit von BERGER, J., BERGMANN, K. E., GREISER, E., KEIL, U., SCHÄFER, H., SCHWARTZ, F. W. (1991): Manual für die Planung und Durchführung epidemiologischer Studien. Schriftenreihe der GMDS (11). Schattauer Verlag, Stuttgart.

WINDELER, J., TRAMPISCH, H. J., DIETLEIN, G., ELZE, M., GÖRTELMEYER, R., HASFORD, J., HAUSCHKE, D., HERBOLD, M., HILGERS, R., LANGE, S. et al. (1995): Empfehlungen zur Durchführung von Studien zur therapeutischen Äquivalenz. Informatik, Biometrie und Epidemiologie in Medizin und Biologie **26** (4), 350–355.

Anschrift der Verfasser: Prof. Dr. H. Schäfer Institut für Medizinische Biometrie und Epidemiologie der Philipps-Universität Marburg Bunsenstraße 3 D-35037 Marburg Tel.: 06421/286207 Fax: 06421/288921

BUCHBESPRECHUNGEN/BOOKREVIEWS

BORTZ, L.
Kurzgefaßte Statistik für die klinische Forschung.
Ein praktischer Leitfaden für die Analyse kleiner Stichproben.
Springer Verlag Berlin, 1998, 406 Seiten, DM 68,–, ISBN 3-540-63738-9

Das zu Anfang des Jahres 1998 erschienene Buch ist ein ebenso Anwender-orientiertes wie -freundliches Extrakt des Klassikers „Verteilungsfreie Methoden in der Biostatistik", das schon aufgrund seines gewaltigen Umfangs von mehr als 2000 Seiten in der letzten Ausgabe nicht mehr zur Standardausstattung des biometrisch interessierten Naturwissenschaftlers oder Mediziners gehören konnte. Bereits beim ersten Blättern wird deutlich, daß man es bei der „Kurzgefaßten Statistik" allerdings nicht mit der k + 1-ten Wiederauflage eines allgemeinen Lehrbuchs der Biometrie zu tun hat: Behandelt werden ausschließlich verteilungsfreie Methoden, auch wenn der Titel jeden Hinweis darauf verschweigt – befürchten die Autoren (vielleicht zu Recht?) eine abschreckende Wirkung durch Worte wie „nonparametrisch" oder „Verteilung" auf dem Buchdeckel?
Im ersten Kapitel, und hier ist die Übereinstimmung mit den klassischen Lehrbüchern noch am ehesten erkennbar, wird eine allgemeine Einführung in die Begriffswelt der Wahrscheinlichkeit und die konfirmatorische Statistik gegeben, wobei allerdings schon an dieser frühen Stelle auf Aspekte aufmerksam gemacht wird, die später das Verständnis der Rangstatistiken erleichtern. Auch auf die Unterscheidung zwischen asymptotischen und exakten Tests wird bereits eingegangen, ebenso auf das Problem „Signifikanztest" versus „klinische Relevanz". Im zweiten Kapitel werden Tests für Häufigkeiten dargestellt, wobei auch Verfahren beschrieben werden, die in herkömmlichen Büchern dieser Art und vor allem dieses Umfangs kaum besprochen werden, z. B. Cochrans Q-Test und der Marginalhomogenitätstest von Lehmacher. Kapitel 3 ist den Rangtests für unabhängige und abhängige Stichproben gewidmet und behandelt auch die sonst wenig beschriebenen, aber oft nützlichen Trendtests, z. B. von Jonckheere oder Page. Kapitel 4 (für das man vielleicht eine etwas griffigere Überschrift als „Testmethoden für Meßwerte" finden könnte) umfaßt die Randomisierungstests und die Familie der Kolmogoroff-Smirnov-Tests. Den Zusammenhangsmaßen ist Kapitel 5 gewidmet; auch hier finden sich wieder Verfahren mit interessanten Anwendungsgebieten, die sonst kaum Erwähnung finden (für wen gehört beispielsweise die „Zwillingskorrelation von Whitfield" zum täglichen Handwerkszeug?). Übereinstimmungsmaße bei subjektiven Merkmalsbeurteilungen sind das Thema in Kapitel 6. In Kapitel 7 werden (verteilungsfreie) Sequentialtests besprochen, die (verteilungsfreie) Analyse von Zeitreihen ist Gegenstand von Kapitel 8, und auch hier wird wieder nahezu Neuland betreten mit einem „Sprungstellen-Detektionstest" oder einem „Häufungstrendtest". Schließlich wird der Appetit auf weiterführende Methoden durch ein – gänzlich formelfreies – Kapitel 9 angeregt.
Der Untertitel mit seinem Verweis auf kleine Stichproben mag Bedenken aufkommen lassen, ob denn nun gerade hier die verteilungsfreien Methoden erbarmungslos mit ihrer Asymptotik eingeführt werden. Nun, man kann beruhigt sein: Wie in kaum einem anderen Buch dieser Art wird bei jedem Verfahren beschrieben, wie man bei kleinen Stichprobenumfängen vorzugehen hat, wobei natürlich oftmals ein Verweis auf den Tabellenanhang erfolgt. Dies mag für manchen ein unangenehmer Weg sein, aber er ist immerhin noch besser gangbar, als wenn man den Kliniker anweisen würde, Permutationstests, shift- oder network-Algorithmen auf seine Daten loszulassen. So wird denn auch stets anhand „echter Zahlen" entschieden, wann eine Stichprobe noch „klein", oder wann Sie schon „groß" ist. Darüber mag man streiten, hilfreich ist es allemal.
Sicher, einige der Verfahren aus diesem Buch sind weitab vom täglichen Gebrauch, aber es ist auch sicher nicht Absicht der Autoren gewesen, ein Nachschlagewerk der Standardverfahren zu präsentieren, denn davon sind bereits etliche auf dem Markt. Dafür ist ihnen das Kunststück gelungen, bei handlichem Umfang ein an Beispielen mit detailliert und sachkundig beschriebenen medizinischen Hintergründen (!) reiches Handbuch zu verfassen, das auch bei ausgefalleneren Problemen sicher ein adäquates Verfahren anbieten kann. Der Theoretiker wird bei Fragen der mathematischen Herleitung der Verfahren auf die entsprechenden Stellen in den „Verteilungsfreien Verfahren" verwiesen, wobei auch stets die betreffenden Originalarbeiten zitiert werden – allein das Literaturverzeichnis ist daher eine wahre Fundgrube. So bleibt für den Anwender der Blick frei für die praktischen

Aspekte, und da jede Methode, jeder Test, jede Kenngröße an klinischen Beispielen erläutert und durchgerechnet wird (wobei diese Abschnitte schon durch den Druck deutlich hervorgehoben sind), dürften kaum Fragen offen bleiben. Der gehobenen Bedeutung des berühmten kleinen p in der klinischen Forschung ist dann wohl auch die Tatsache zuzuschreiben, daß es in diesem Buch stets groß geschrieben wird ... Thomas Bregenzer, Berlin

ORTSEIFEN, C.
Der SAS Kurs. Eine leicht verständliche Einführung
Bonn: International Thomson Publishing 1997
316 S.; 1. Auflage; kt. 59, DM; ISBN 3-929821-44-3

Das Buch gibt eine deutschsprachige Einführung in das SAS System. Es wird begleitet von einer 3,5"-Diskette, die zu Übungszwecken SAS-Programme, SAS-Dateien sowie kommentierende Texte enthält. Der Umfang sowie der Preis des Buches ist als Einführungstext akzeptabel.
Das Inhalts- und Stichwortverzeichnis ermöglichen das rasche Auffinden spezieller Aspekte. Der Inhalt des Buches gliedert sich in 11 Kapitel und 6 Anhänge (Kap. 1 Zum Gebrauch des Buches, Kap. 2 Der modulare Aufbau des SAS-Systems, Kap. 3 SAS-Programme und SAS-Dateien, Kap. 4 Der Display Manager, Kap. 5 Dateien anlegen, Kap. 6 Daten in Tabellenform präsentieren, Kap. 7 Dateien bearbeiten, Kap. 8 Statistische Analysen, Kap. 9 Grafik, Kap. 10 Optionen, Titel, Tools und Keys, Kap. 11 Ein Ausblick). Kleinere Aufgaben am Ende der Kapitel 3 bis 10 unterstützen den tutoriellen Charakter des Buches. Alle Erklärungen basieren auf der im Moment aktuellen Version SAS 6.12. Die wesentlichen Aspekten werden im Text anhand von Beispielprogrammen, SAS-Outputs sowie Protokollen des Protokollfensters (LOG) illustriert.
Verschiedene Symbole am Textrand geben dem Leser eine schnelle Orientierung zu wesentlichen Aspekten im Text, zu Beispielprogrammen, die auf der Diskette mitgeliefert werden, zur allgemeinen Syntaxbeschreibung und zu Übungen für den Leser.
Die Schwerpunkte des Buches liegen in der Erstellung von SAS-Dateien, der Datenanalyse und Präsentation von Ergebnissen. Wesentliche Dinge, die gerade für den Einsteiger in SAS wertvoll sind, werden gut strukturiert dargestellt. Zum Beispiel wird das Konzept der temporären und permanenten SAS-Datei sowie die Reihenfolge von Programmabläufen in SAS verständlich erklärt. Für die maskengesteuerte Eingabe in SAS wird PROC FSEDIT erläutert. Die wesentlichen analytischen Prozeduren (PROC FREQ, MEANS, UNIVARIATE, CORR, TTEST, NPAR1WAY) werden erklärt. Es werden auch manche wertvolle Tips gegeben, die sicherlich auch erfahreneren SAS-Programmierern nicht bekannt sind.
Besonders hervorzuheben ist Kapitel 9. Es gibt eine gute Einführung bezüglich der Erzeugung von visuell ansprechenden SAS-Grafiken mit Hilfe von PROC GPLOT und erspart gerade dem Einsteiger das unnötige Erstellen von Grafiken in SAS-fremden Programmen (z.B. Excess, Access usw.). Darüberhinaus vermittelt das Buch auch Techniken, wie mittels SAS erzeugte Grafiken exportiert bzw. in Textdokumente importiert werden können.
Das Buch ist auch in der häufigen Situation eine Hilfe, in der SAS-fremde Dateien in SAS eingelesen werden sollen. Hierzu werden im Kapitel 11 wichtige Prozeduren bzw. interaktive Lösungen mit Hilfe des Import Wizards geboten.
Der Autorin Carina Ortseifen ist es gelungen, ein ausgewogenes Buch für SAS-Einsteiger zu schreiben. Es kann meines Erachtens guten Gewissens Einsteigern, insbesondere auch medizinischen Doktoranden, empfohlen werden.

Verfasser:
Dr. med. Andreas Stang, MPH(USA)
Institut für Medizinische Informatik, Biometrie und Epidemiologie
Universitätsklinikum Essen
Hufelandstr. 55
45122 Essen

Tel.: 0201-723-4524
Fax: 0201-723-5701
e-mail: andreas.stang@uni-essen.de

Mitteilungen der
Deutschen Gesellschaft für Medizinische Informatik,
Biometrie und Epidemiologie (GMDS) e.V.

Inhalt nach Rubriken

Brief des Präsidenten
Prof. Dr. K.-H. Jöckel

Liebe Kolleginnen und Kollegen,

inzwischen ist Ihnen das vorläufige Programm der 44. Jahrestagung unserer Gesellschaft vom 13. - 16.09.1999 in Heidelberg zugegangen. Die gemeinsame Veranstaltung mit der International Society for Clinical Biostatistics, die sich diesjährig zum 20. mal trifft, eröffnet unseren biometrisch interessierten Kolleginnen und Kollegen Möglichkeiten eines Ideen- und Erfahrungsaustausches mit Experten aus aller Welt. Aber auch für die Epidemiologie und die Medizinische Informatik wurde ein attraktives Programm zusammengestellt. Den Tagungsleitern, Professor Victor, Professor Haux, Professor Wahrendorf und Dr. Edler den Vorsitzenden und Mitgliedern der beiden Programmkomitees sowie dem lokalen Organisationskomitee bereits heute unseren herzlichen Dank! Ich hoffe, Sie werden zahlreich in Heidelberg erscheinen und eine erfolgreiche Tagung erleben.

Aus der Geschäftsstelle gibt es zu berichten, daß Herr Dipl.-Vw. Thomas Banasiewicz uns verlassen hat, um eine Stelle in der freien Wirtschaft anzunehmen. Herr Banasiewicz hat in den etwas mehr als 2 Jahren, in denen er unserer Gesellschaft als Geschäftsführer und zuletzt auch als Schatzmeister gedient hat, viel für die Konsolidierung der Geschäftsstelle und die Leistungsfähigkeit der Gesellschaft getan. Ich darf ihm auch auf diesem Wege für die geleistete Arbeit danken und ihm für sein weiteren Lebensweg, in dem er sicherlich der Gesellschaft verbunden bleiben wird, alles Gute wünschen. Als Nachfolgerin in der Geschäftsstelle kann ich Ihnen Frau Friederike Sträter vorstellen, die mit der einen Hälfte ihrer Stelle die Geschäftsstelle der GMDS, mit der anderen die der Deutschen Gesellschaft für Regulatorische Angelegenheiten e. V. (DGRA) leitet. Ich darf Sie bitten, Frau Sträter nach Kräften bei ihrer Arbeit zu unterstützen. Wie Sie sich sicherlich denken können, kann eine solche Personalreduktion nicht ohne Folgen für das Leistungsspektrum der Geschäftsstelle bleiben. Wir bitten Sie also bereits heute um Verständnis, sollte es in einzelnen Situationen zu Schwierigkeiten und Engpässen kommen. Dem Präsidium ist klar, daß eine besondere Unterstützung durch die Mitglieder des Präsidiums für die Geschäftsstelle in dieser Situation dringend erforderlich ist. Über die zukünftige weitere Ausgestaltung der Geschäftsstelle wird das Präsidium demnächst beraten.

Dieses wird und kann nur unter Würdigung der finanziellen Rahmenbedingungen geschehen.

Teil dieser finanziellen Rahmenbedingungen sind die nicht unerheblichen finanziellen Mittel, die die Gesellschaft jährlich für Ihre Publikationen aufwendet. Aus diesem Grunde hat sich das Präsidium entschlossen, vorsorglich den Vertrag mit den Verlagen Eugen Ulmer, Urban & Fischer, die den "Silberfisch" herausgeben, zu kündigen. Derzeit führen wir Verhandlungen mit anderen Verlagen, um hier zu einer kostengünstigeren Lösung zu kommen. Darüber hinaus werden wir uns bemühen, durch eine straffere Führung und Haushaltspolitik bei der Organisation von Jahrestagungen Einsparungspotentiale, z. B. durch eine geschickte Auftragsvergabe zu realisieren. Hier werden wir auf Vorarbeiten von Herrn Banasiewicz und die Erfahrungen von Frau Sträter zurückgreifen können.

Nachdem wir mit der Verankerung der Sektion "Medizinische Dokumentation" in der Satzung unserer Gesellschaft für Medizinische Dokumentare geöffnet haben, hat das Präsidium auf seiner letzten Sitzung der Einrichtung einer Vorkommission zum Zertifikat "Medizinischer Dokumentar" zugestimmt. Aufgabe dieser Kommission ist es Zertifikatsrichtlinien zu erarbeiten, die eine MD Zertifizierung von Berufstätigen mit MDA-Abschluß ermöglicht.

Dieser Vorkommission gehören an: auf Vorschlag des DVMD Frau Walter-Jung und Herr Linczak, auf Vorschlag der GMDS Herr Professor Bernauer und Herr Professor Gaus, als von beiden Organisationen benannter Vertreter und Vorsitzender ich selber.

Vom Zertifikat für Biometrie gibt es zu berichten, daß ein Schriftwechsel mit der EMIA vorliegt, daß seitens der EMIA kein Zweifel an der mit dem Zertifikat nachgewiesenen Qualifikation im Hinblick auf die Funktion eines verantwortlichen Biometrikers in einer klinischen Prüfung gemäß ICH-Guidelines besteht. Das Zertifikat Epidemiologie hat ebenfalls noch einmal an Bedeutung dadurch gewonnen, daß es als hinreichende Bedingung für den Nachweis epidemiologischer Qualifikation von Anforderern anonymisierter Daten aus bevölkerungsbezogenen Krebsregistern anerkannt ist.

Ich möchte Ihnen noch das Ergebnis der Wahlen zum Präsidium mitteilen: Herr Professor Lehmacher ist mit 173 Stimmen von 313 gültig abgegebenen Stimmen zum 1. Vizepräsidenten gewählt worden. Dr. Zaiß erhielt 270 Stimmen als neuer Schriftführer und Professor Schweim 266 Stimmen als neuer Schatzmeister der GMDS. Die Auszählung erfolgte durch Dr. Stausberg in Anwesenheit des GMDS-Mitgliedes Frau Katja Bromen. Die Gewählten haben die Wahl angenommen und ich darf ihnen auf diesem Wege herzlichst gratulieren und ihnen eine erfolgreiche Arbeit wünsche.

Am Schluß meiner Amtszeit möchte ich allen, die mich in meiner Arbeit als Präsident unterstützt haben, herzlich danken. Besonderer Dank gilt den Kolleginnen und Kollegen aus dem Präsidium, insbesondere meinem Vorgänger, Herrn Professor Köpcke, der aus dem Präsidium ausscheidet, und meinem Nachfolger, Herrn Professor Klar, dem ich eine glückliche Hand bei der weiteren Lenkung des Schicksals unserer Fachgesellschaft wünsche. Mein Dank gilt aber auch den Leitern und Mitgliedern der Arbeitskreise, Projekt- und Arbeitsgruppen und der Kommissionen. Stellvertretend für viele möchte ich Herrn Professor Schäfer, Marburg, den Leiter der Präsidiumskommission Arzneimittelgesetz und Arzneimittelprüfrichtlinien und langjähriges Beiratsmitglied nennen, der die Arbeit des Präsidenten im Hinblick auf die europäischen Entwicklungen in seinem Aufgabengebiet beispielhaft unterstützt hat. Last but not least gilt mein Dank meinem Mitarbeiter Herrn Dr. Stausberg, der als Schriftführer unermüdlich seinen Dienst getan hat.

Für das dritte Quartal des Jahres gratuliere ich den Kollegen Herrn Professor Kres, Herrn Dr. Lajosi, Herrn Professor Scheibe und Frau Dr. Török zum Geburtstag , im Namen der GMDS, recht herzlich.

In der Hoffnung, Sie möglichst zahlreich in Heidelberg begrüßen zu können, verbleibe ich

mit freundlichen Grüßen
Ihr

Prof. Dr. K.-H. Jöckel
Präsident

Aus-, Fort- und Weiterbildung

Veranstaltungen der Akademie Medizinische Informatik

1. Tutorium **"Medizin-Controlling und dessen Unterstützung durch Informationstechnologie"**

Termin, Ort: 12.9.99, im Rahmen der internationalen Tagung ISCB-GMDS-99, Heidelberg

Zielgruppe: Ärzte, Medizininformatiker, Pflegekräfte und Medizinische Dokumentare, die im Bereich des Medizin-Controlling tätig sind oder damit zu tun haben bzw. beruflich planen. Interessant ist das Tutorium auch für Mitarbeiter der Medizinischen Informatik im allgemeinen, im ärztlichen und Pflegedienstbereich, im Patientenmanagement sowie im Controlling, der Kosten-Leistungsrechnung und innerhalb der Krankenhaus-Betriebsleitung.

Referent: Univ.-Prof. Dr. Andreas J. W. Goldschmidt, Universität Bonn

Teilnahmegebühren: Mitglieder von GI, BVMI, DVMD, **GMDS 300.- DM**
(Sonstige 450.- DM, Studierende 200.- DM)

Hinweis: *Dieses Seminar kann für den Erwerb des gmds-Zertifikats 'Medizinische Informatik' angerechnet werden.*

2. Seminar **"Kodierungssoftware für Diagnosen und Prozeduren im Vergleich"**

Termin, Ort: 17.9.99, Kopfklinik, Heidelberg

Zielgruppe: Anbieter und Anwender von KIS-Systemen

Referenten: Dr. Josef Ingenerf, Universität Lübeck; PD Dr. med. Roswitha Thurmayr, TU München

Teilnahmegebühren: Mitglieder von BVMI, DVMD, **GMDS 320.- DM**
(Sonstige 800.- DM, Studierende 160.- DM)

Hinweis: *Dieses Seminar kann für den Erwerb des gmds-Zertifikats 'Medizinische Informatik' angerechnet werden.*

3. Seminar **"Analyse, Auswahl und Einführung von Krankenhausinformationssystemkomponenten"**

Termin, Ort: 7.10.99, Leipzig

Zielgruppe: Das Seminar wendet sich an Personen in Krankenhäusern, die mit der Einführung von Komponenten für Krankenhausinformationssysteme befaßt sind. Hierzu gehören insbesondere Abteilungsleiter, Projektleiter oder alle Personen, die Kenntnisse über das taktische Management von Krankenhausinformationssystemen erwerben oder ihr bisheriges Wissen aktualisieren wollen. Darüber hinaus richtet sich das Seminar auch an interessierte Personen bei Beratungsunternehmen bzw. Anbietern von Komponenten für Krankenhausinformationssysteme.

Referenten: Prof. Dr. sc. hum. Alfred Winter, Dipl.-Math. Gabriele Herrmann, Dr. oec. habil. Alfred Scharsky, Dr. sc. hum. Birgit Schneider, Universität Leipzig

Teilnahmegebühren: Mitglieder von BVMI, DVMD, **GMDS 320.- DM**
(Sonstige 800.- DM, Studierende 160.- DM)

Hinweis: *Dieses Seminar kann für den Erwerb des gmds-Zertifikats 'Medizinische Informatik' angerechnet werden.*

4. Seminar **"Einführung von Komponenten für Klinische Arbeitsplatzsysteme - Probleme, Hemmnisse und Lösungsstrategien"**

Termin, Ort: 8.10.99, Universität Leipzig

Zielgruppe: Das Seminar wendet sich an Personen in Krankenhäusern, die mit der Planung, Einführung und Betreuung von Komponenten für Klinische Arbeitsplatzsysteme beauftragt sind. Hierzu gehören insbesondere Medizininformatiker, Informatiker, Ärzte, Projektleiter im Rahmen von Einführungsprozessen, Mitarbeiter des Pflegedienstes etc. Darüber hinaus richtet sich das Seminar auch an interessierte Personen bei Beatungsunternehmen bzw. Anbietern von Komponenten für Krankenhausinformationssysteme.

Referentin: Dipl.-Math. Gabriele Herrmann, Universität Leipzig

Teilnahmegebühren: Mitglieder von BVMI, DVMD, **GMDS 320.- DM**
(Sonstige 800.- DM, Studierende 160.- DM)

5. Seminar **"Gesundheitsökonomie"**
Termin, Ort: 11.-12.10.99, DKFZ, Heidelberg
Zielgruppe: Personen, die in die Planung und Durchführung gesundheitsökonomischer Studien involviert bzw. an Entscheidungsprozessen beteiligt sind, und die das Ziel einer (verbesserten) wirtschaftlichen Gesundheitsversorgung von Patienten verfolgen.
Personen, die an der Erfassung, Verwaltung und Verarbeitung von Kosten- und Leistungsdaten für Controlling und zum Qualitätsmanagement arbeiten.
Ärzte, die bei der Therapiewahl auch ökonomische Informationen bewerten bzw. einbeziehen möchten.
Personen, die mehr über Kosten- und Nutzenbewertung im Gesundheitswesen wissen möchten.
Spezifische Vorkenntnisse werden nicht vorausgesetzt.
Referent: Dipl.-Inform. Med. Oliver Mast, Leverkusen
Teilnahmegebühren: Mitglieder von BVMI, DVMD, **GMDS 500.- DM**
(Sonstige 1.200.- DM, Studierende 300.- DM)
Hinweis: *Dieses Seminar kann für den Erwerb des gmds-Zertifikats 'Medizinische Informatik' angerechnet werden.*

6. Seminar **"Verschlüsselung von Diagnosen und Operationen mit ICD-9/10 und OPS"**
Termin, Ort: 13.10.99, DKFZ, Heidelberg
Zielgruppe: Ärzte, Medizinische Dokumentare, Medizinische Informatiker
Referent: Dr. med. Albrecht Zaiss, Universität Freiburg
Teilnahmegebühren: Mitglieder von BVMI, DVMD, **GMDS 320.- DM**
(Sonstige 800.- DM, Studierende 160.- DM
Hinweis: *Dieses Seminar kann für den Erwerb des gmds-Zertifikats 'Medizinische Informatik' angerechnet werden.*

7. Seminar **"Medizin für Nicht-Mediziner"**
Termin, Ort: 14.-15.10.99, RWTH Aachen
Zielgruppe: Informatiker, Medizininformatiker, Studierende der Medizinischen Informatik und der Informatik
Referent: Prof. Dr. Dr. Klaus Spitzer, RWTH Aachen
Teilnahmegebühren: Mitglieder von BVMI, DVMD, **GMDS 500.- DM**
(Sonstige 1.200.- DM, Studierende 300.- DM
Hinweis: *Dieses Seminar kann für den Erwerb des gmds-Zertifikats 'Medizinische Informatik' angerechnet werden.*

Postgraduelle Ausbildung MEDIZINISCHE BIOMETRIE

Die Medizinische Fakultät der Universität Heidelberg bietet, unter Federführung des Instituts für Medizinische Biometrie und Informatik, ein postgraduelles Ausbildungsprogramm in **MEDIZINISCHER BIOMETRIE** an. Die Ausbildung kann flexibel in thematisch abgeschlossenen Blöcken mit individueller Kurswahl absolviert werden, so daß eine berufsbegleitende Weiterbildung möglich ist. Bei erfolgreichem Durchlaufen eines definierten Curriculums führt das Programm zu einem Universitätszertifikat mit Fachanerkennung der GMDS.

Die Kurse können auch einzeln besucht werden, ohne daß die Erlangung des Zertifikats angestrebt wird.

In der zweiten Jahreshälfte 1999 werden die folgenden Kurse angeboten. Dabei werden die genannten Kursleiter durch weitere Dozenten unterstützt.

GRUNDLAGENKURS MEDIZIN

08.10. - 10.10.1999	Bewegungsapparat / Orthopädie
	Dr. K.-L. Krämer, Universität Heidelberg

GRUNDLAGENKURS STATISTIK

Prof. Dr. F.-T. Nürnberg, FH Mannheim

19.08. - 23.08.1999, 02.09. - 06.09.1999, 09.09. - 12.09.1999

AUFBAUKURSE

11.11. – 13.11.1999	Überlebenszeitanalyse
	Prof. Dr. M. Schumacher, Universität Freiburg
25.11. – 27.11.1999	Klinische Studien, Teil II
	Dr. J. König, Universität Homburg/Saar

Wahlkurse

26.08. – 28.08.1999	Bayes-Methoden in der Medizin
	Dr. K. Ickstadt, TU Darmstadt
30.09. – 02.10.1999	Nichtparametrische Verfahren in der Biometrie
	Prof. Dr. E. Brunner, Universität Göttingen
02.12. – 04.12.1999	Zeitreihenanalyse und Kurvenschätzung
	Prof. Dr. R. Dahlhaus, Universität Heidelberg

Weitere Kurse sind in Planung. Die Kurse sind stark anwendungsorientiert und die Lehrinhalte werden mittels praxisnaher Computerübungen vertieft.

Für weitergehende Informationen wenden Sie sich bitte an:

Dr. Katrin Jensen, Abteilung Medizinische Biometrie der Universität Heidelberg
Im Neuenheimer Feld 305, 69120 Heidelberg
Tel: 06221/56-4180, -4141; FAX: 06221/56-4195; E-Mail: jensen@imbi.uni-heidelberg.de

Dr. Birgit Stadler, Zentrum für Studienberatung und Weiterbildung
Friedrich-Ebert-Anlage 22-24, 69117 Heidelberg
Tel: 06221/54-7815, -7810; FAX: 06221/54-7819, E-Mail: Birgit.Stadler@urz.uni-heidelberg.de
http://www.rzuser.uni-heidelberg.de/~ah5/postgrad.html

Veranstaltungen des Zentrums Biometrie 1999

Veranstaltung:	**SAS-Kurs für Anfänger***
Termine, Ort:	02. - 04.09.1999 in der Ruhr-Universität Bochum
Zielgruppe:	Mediziner, Statistiker, Medizinische Dokumentare und andere Personen, die das selbständige Progammieren mit SAS erlernen wollen.
Referent:	Heinrich Stürzl, Marburg
Teilnahmegebühren:	Mitglieder: 900,-DM, Studierende 150,-DM, Sonstige 1200,-DM

Veranstaltung:	**Einführung in Relationale Datenbanken - SQL**
Termin, Ort:	23. - 25.09.1999, in der Ruhr-Universität Bochum
Zielgruppe:	Personen, die Kenntnisse über den Entwurf relationaler Datenbanken, den Aufbau von Tabellen sowie die Datenabfrage über SQL haben müssen
Referentin:	Dipl.- Dok. Susanne Stolpe, Bochum
Teilnahmegebühren:	Mitglieder: 900,-DM, Studierende 150,-DM, Sonstige 1200,-DM

Veranstaltung:	**PL/SQL (Oracle)**
Termin, Ort:	14. - 15.10.1999, in der Ruhr-Universität Bochum
Zielgruppe:	Der Kurs PL/SQL richtet sich an Personen, die mit dem relationalen Datenbank-System Oracle 7.3 arbeiten und die z.B. im Rahmen der Programmierung von Triggern, Plausibilitätskontrollen oder komplexen Prozeduren die im Vergleich zu SQL erheblich größeren Möglichkeiten der Oracle Programmiersprache PL/SQL kennenlernen wollen. SQL-Kenntnisse werden vorausgesetzt
Referentin:	Dipl.- Dok. Susanne Stolpe, Bochum
Teilnahmegebühren:	Mitglieder: 600,-DM, Studierende 100,-DM, Sonstige 700,-DM

Veranstaltung:	**Weiterführende Statistik**
Termin, Ort:	21. - 22.10.1999, in der Ruhr-Universität Bochum
Zielgruppe:	Das Seminar richtet sich an Personen, die medizinische Daten auswerten und dieses nicht nur auf der Basis von Zwei-Gruppen-Vergleichen durchführen wollen.
Referentin:	Dipl.-Stat. Martina Kron, Universität Ulm
Teilnahmegebühren:	Mitglieder: 600,-DM, Studierende 100,-DM, Sonstige 700,-DM

Veranstaltung:	**SAS-Kurs für Fortgeschrittene***
Termine, Ort:	28. - 30.10.1999 in der Ruhr-Universität Bochum
Zielgruppe:	Mediziner, Statistiker, Medizinische Dokumentare
Referent:	Heinrich Stürzl, Marburg
Teilnahmegebühren:	Mitglieder: 900,-DM, Studierende 150,-DM, Sonstige 1200,-DM

Veranstaltung:	**Einführung in die Biometrie**
Termin, Ort:	18. - 19.11.1999, in der Ruhr-Universität Bochum
Zielgruppe:	Personen ohne oder mit geringen statistischen Kenntnissen, die biometrische Verfahren verstehen oder selbst anwenden wollen.
Referentin:	Ingrid Spreckelsen, Bochum
Teilnahmegebühren:	Mitglieder: 600,-DM, Studierende 100,-DM, Sonstige 700,-DM

Veranstaltung:	**Medizin für Nichtmediziner**
Termin, Ort:	25. - 27.11.1999 in der Ruhr-Universität Bochum
Zielgruppe:	in der medizinischen Biometrie tätige, die an der Planung, Durchführung und Auswertung von Studien beteiligt sind und wenig medizinische Vorbildung haben
Referentin:	Dr. Claudia Hänel, Düsseldorf
Teilnahmegebühren:	Mitglieder: 900,-DM, Studierende 150,-DM, Sonstige 1200,-DM

(* in Kooperation mit der postgraduellen Ausbildung „Medizinische Biometrie" der Universität Heidelberg. Für dieTeilnehmer dieser Ausbildung gelten die Heidelberger Kurspreise)

Auskunft: Walter Dieckmann
Akademie für öffentliche Gesundheit
44780 Bochum
tel: 0234 7005162 fax 0234 7094325
email: Walter.Dieckmann@ruhr-uni-bochum.de
www.biometrie.net

Ankündigungen und Veranstaltungshinweise

MIE 1999 Bridges to the Knowledge

22.8. – 26.8. 1999 in Ljubljana, Slowenien
Weitere Informationen unter http://www.mie99.org/Home1.htm

---

44. Jahrestagung der GMDS
(Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie)
13.-16. September 1999 in Heidelberg

Kontakt: ISCB-GMDS-99 Conference Office
c/o Dept. of Medical Biometry
University of Heidelberg
Im Neuenheimer Feld 305
D-69120 Heidelberg, Germany
Phone: ++ 49 6221 56 - 5656
Fax: ++ 49 6221 56 - 4195

Weitere Informationen unter http://www.dkfz-heidelberg.de/biostatistics/iscb-gmds-99/

MEDNET '99: 4. WORLD CONGRESS ON THE INTERNET IN MEDICINE
18.-21.September 1999 in Heidelberg

Ansprechpartner:
Universität Heidelberg; Institut für Klinische Sozialmedizin
= MEDNET99 Congress Secretariat =
Bergheimerstr. 58
69115 Heidelberg
Germany

Ausführliche und aktuelle Informationen jetzt über http://yi.com/mednet99/ erhältlich!

KI '99
23. Deutsche Jahrestagung für Künstliche Intelligenz
13. - 15. September 1999 in Bonn
Weitere Informationen unter http://wob.informatik.uni-bonn.de/Wob/de/view/ki99.html

Informatik '99
29. Jahrestagung der GI
5. – 9. Oktober 1999 in Paderborn
Weitere Informationen unter http://www.uni-paderborn.de/cs/informatik99/

---

4. Forum für Information im Krankenhaus
- Krankenhaus Online -
18. - 22. Oktober 1999
Neue Messe München, Halle A5

Sonderschau, Symposium und Seminar

Die Sonderschau „K online" zeigt eine Auswahl beispielhafter Telematiklösungen aus Forschung und Industrie für den Einsatz in der täglichen Praxis

Das Symposium zeigt aktuelle Trends und Problemlösungen der Informationsverarbeitung in der Medizin gerade zum Wechsel und für das Jahr 2000 und folgende

Das Seminar bringt Grundlagen moderner Informationssysteme in Krankenhaus und Arztpraxis.

Unterstützt durch PROREC (TAP) , BVMI und GMDS

Weitere Informationen:
Dr. Rolf Engelbrecht
GSF-MEDIS
Ingolstädter Landstr. 1
85764 Neuherberg

Fax : +49-89-3187-3008
E-mail: engel@gsf.de
Internet: www-mi.gsf.de/konline

Telemed '99
5.-6. November 1999 in Berlin
Arbeitsgruppe Telemedizin der GMDS, Freie Universität Berlin und BVMI
Nach den erfolgreichen Tagungen im November 1996, 1997 und 1998 veranstaltet die Landesvertretung Berlin/Brandenburg des BVMI in Zusammenarbeit mit der Freien Universität Berlin und der Projektgruppe Telemedizin der GMDS nun die 4. Fortbildungsveranstaltung und Arbeitstagung zu der immer noch hochaktuellen Thematik Telemedizin - Telematikanwendungen im Gesundheitswesen. Zielsetzung der Veranstalter ist, auf der TELEMED'99 neue Ergebnisse telemedizinischer Anwendungen zur Diskussion zu stellen, Erfahrungen auszutauschen und eine Plattform zur Koordinierung der vielen Aktivitäten auf dem Gebiet der Telematik im Gesundheitswesen zu bilden, insbesondere für die Telemedizin und im Bereich medizinischer Netze.
Weitere Informationen unter http://www.medizin.fu-berlin.de/medinf/telemed99/

HIGH CARE 2000
25. – 27. Februar 2000 in Bochum,
Veranstalter: Prof. Dr. Dietrich Grönemeyer, M.D, Department of Radiology and MicroTherapy, University of Witten/Herdecke, Germany
Kontakt: Angela Hollmann
Phone: +49 (0) 234 9780 - 114
Fax: +49 (0) 234 9780 - 599
E-mail: hollmann@microtherapy.de
weitere Informationen unter http://www.highcare.de/

Bildverarbeitung für die Medizin 2000
12. - 14. März 2000 in München
Kontakt: PD Dr. Dr. Alexander Horsch
Institut für Medizinische Statistik und Epidemiologie
Klinikum rechts der Isar der TU München
Ismaninger Str. 22
81675 München
Tel.: 089 / 4140 4330
Fax.: 089 / 4140 4974
Alexander.horsch@imse.med.tu-muenchen.de
Weitere Informationen: http://www.imse.med.tu-muenchen.de/mi/bvm2000/

Praxis der Informationsverarbeitung im Krankenhaus (KIS-Tagung 2000)
06.-07.04.2000 in Frankfurt a.M./Offenbach

Kontakt: Prof. Dr. Wolfgang Giere — Organisationsleitung
Zentrum der Medizinischen Informatik (Zinfo)
Johann Wolfgang Goethe-Universität
Theodor-Stern-Kai 7
60590 Frankfurt/Main
Tel: 069-6301-6640 (Fr. Galonska)
Fax: 069-6301-6777
email: KIS2000@merlin.add.uni-frankfurt.de

Prof. Dr. Andreas Goldschmidt — Organisationsleitung
Institut für Medizinische Statistik, Dokumentation und Datenverarbeitung (IMSDD)
Rheinische Friedrich Wilhelm-Universität
Sigmund-Freud-Straße 25
53105 Bonn
Tel: 0228-287-4301 (Fr. Felten)
Fax: 0228-287-5032
email: KIS2000@imsdd.meb.uni-bonn.de

Medical Informatics Europe 2000 (MIE 2000) / GMDS 200
27.08.-01.09.2000 in Hannover

Kontakt: Dr. Rolf Engelbrecht
Vorsitzender des Organisationskomitees
GSF- Forschungszentrum für Umwelt und Gesundheit, medis-Institut
Ingolstädter Landstr. 1
85764 Neuherberg
Tel: +49-89-3187-4138
Fax: +49-89-3187-3008
email: engel@gsf.de

Prof. Dr. Joachim Dudeck
Vorsitzender des Organisationskomitees
Justus-Liebig-Universität, Institut für Med. Informatik
Heinrich-Buff-Ring 44
35392 Gießen
Tel: +49-641-99-413 50
Fax: +49-641-99-413 59
email: joachim.dudeck@informatik.med.uni-giessen.de

weitere Informationen unter http://www-mi.gsf.de/mie2000/

Informatik 2000
30. Jahrestagung der GI
20. – 23. September 2000 in Berlin

Neuaufnahmen

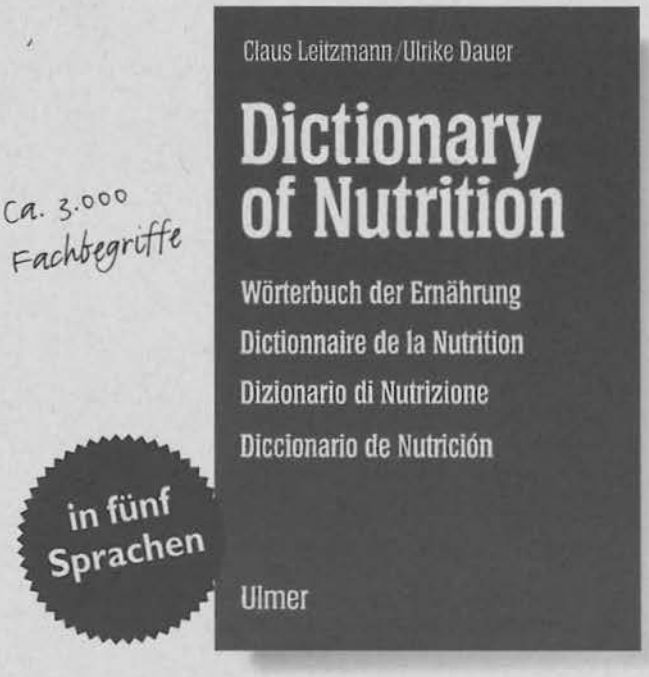
Als neue Mitglieder begrüßen wir recht herzlich

Name/Vorname/Akad. Titel	Ort	Name/Vorname/Akad. Titel	Ort
Alzinger, Johann.	Poing	Kugler, Joachim, Prof. Dr. med.	Dresden
Baumgardt-Elms, Cornelia	Hamburg	Lerch, Magnus	Braunschweig
Boeker, Martin, Dr. med.	Freiburg	Ludwig, Annett	Saarburg
Brenke, Elke, Dr. med.	Köln	Lüdtke, Irene, Dr. med.	Köln
Cnota, Peter	Flörsheim	Meyers, Heribert, Dr. med.	Ober-Ramstadt
Eichner, Eckhard, Dr. med.	Augsburg	Mücke, Wolfgang-Günter	Düsseldorf
Eisinger, Bettina, Dr. med.	Berlin	Nüfer, Michael	Hürth/Efferen
Elbern, Frank	Diusburg	Ockenfels, Ellen, Dr. med.	Köln
Faldum, Andreas, Dr. rer.nat.	Heidesheim	Pschichholz, Andreas, Dr. rer. nat	Freiburg
Gerken, Michael Dr. med.	Neuherberg	Pelikan, Ernst	Freiburg
Grigull, Andreas	Bonn	Rechenberger, Klaus	Düsseldorf
Hajen Harald	Köln	Sommerhäuser, Burkhard	Bonn
Heckelbacher, Burkhard, Dr. agr.	Freising	Spiller, Joachim, Dr. med.	Kassel
Hellmich, Martin, Dr. rer.medic.	Köln	Veniseleas, Dimitros	Bochum
Imhoff, Michael, Dr. med.	Dortmund	Walz, Michael, Dr. med.	Mannheim
Janke, Dietmar, Dr. med.	Braunschweig	Weiß, Uwe	Freiburg
Kirchner, Michael, Dr. med.	Frankenberg	Wolters, Bernd, Dr. med.	Bad Oyenhausen

Notizen

Ein umfangreiches Nachschlagewerk

Die englische Sprache ist in der wissenschaftlichen Kommunikation dominant. Ein Großteil des Austausches findet jedoch in der jeweiligen Landessprache statt. Daher ist es sehr wichtig, die gesuchten Fachbegriffe schnell und problemlos in die wichtigsten europäischen Sprachen übersetzen zu können.



ca. 3.000 Fachbegriffe

in fünf Sprachen

Dictionary of Nutrition.
Wörterbuch der Ernährung. Englisch-Französisch-Deutsch-Italienisch-Spanisch.
Prof. Dr. C. Leitzmann, U. Dauer.
2. Auflage 1996.
516 Seiten.
DM 128,–
öS 934.– / sFr 114.–.
ISBN 3-8001-2148-4

Bisher waren Fachbegriffe der Ernährungsphysiologie oft nur unzureichend in einzelne Sprachen übersetzt und schwer auffindbar. Dieses Buch hilft, die fachbezogene Kommunikation auf internationaler Ebene zu erleichtern und zu fördern.

Zum Buch

Mit diesem Werk liegt ein Verzeichnis von relevanten Begriffen für die Ernährungswissenschaft in fünf wichtigen europäischen Sprachen vor. **DIESES BUCH ENTHÄLT UNGEFÄHR 3000 FACHBEGRIFFE.** Im ersten Teil des Buches werden die englischen Stichwörter in alphabetischer Reihenfolge aufgeführt, in englischer Sprache definiert und in vier Sprachen übersetzt. Der zweite Teil enthält die Übersetzungen aller Stichwörter, in den vier Sprachen getrennt. Mittels einer Verweisnummer ist der entsprechende Begriff im ersten Teil des Buches zu finden.

Aus dem Inhalt

3.000 englische Stichwörter mit Definitionen und Übersetzung in 4 Sprachen. Stichwörter in Deutsch, Französisch, Italienisch, Spanisch.

Der Autor

Prof. Dr. rer. nat. C. Leitzmann arbeitet am Institut für Ernährungswissenschaft der Justus-Liebig-Universität Gießen. Er ist Autor vieler Veröffenlichungen und Mitglied zahlreicher wissenschaftlicher Gesellschaften.

Coupon Ihrer Buchhandlung geben oder senden an: Verlag Eugen Ulmer, Postfach 70 05 61, 70574 Stuttgart. Fax: 0711/4507-120

BUCH-COUPON

☐ Senden Sie mir das Buch **„Dictionary of Nutrition"** zum Preis von DM 128,– / öS 934.– / sFr 114.–. Best.-Nr. 21484.

☐ Senden Sie mir kostenlos Ihr Gesamtverzeichnis „Ulmenblatt"

Datum/Unterschrift

Name/Vorname

Straße/Nr.

PLZ, Ort

Biometrie 3/99

E.U. VERLAG EUGEN ULMER

Hoechst Marion Roussel

Biometrikerin in der klinischen Arzneimittelentwicklung

bei Hoechst Marion Roussel

Hoechst Marion Roussel,
das Pharma-Unternehmen von Hoechst erforscht und entwickelt innovative Arzneimittel für wesentliche Krankheitsgebiete und stellt diese so schnell wie möglich Fachärzten und Patienten weltweit zur Verfügung

Wir suchen für die Abteilung Biometrie und Klinisches Datenmanagement am Standort Frankfurt zum nächstmöglichen Zeitpunkt eine/n

Diplom-Statistiker/in oder
Diplom Mathematiker/in
mit Studienschwerpunkt Statistik.

Ihr Aufgabengebiet liegt in der biometrischen Betreuung (Protokollerstellung, statistische Datenauswertung und Berichterstellung) internationaler klinischer Studien der Phasen I-III und in der Planung und Erstellung studienübergreifender Pharmaregistrierungsunterlagen. Ein anderer besonderer Schwerpunkt Ihrer Tätigkeit ist dabei die frühe klinische Entwicklungsphase, in der in Zusammenarbeit mit klinischen Pharmakologen und anderen Experten Studien zur Optimierung und Charakterisierung neuartiger Wirkstoffe geplant und ausgewertet werden.

Sie sind eingebunden in ein multidisziplinarisches, global operierendes, klinisches Team und arbeiten eng mit den medizinischen und biometrischen Projektkollegen aus USA zusammen.

Worauf es uns ankommt:

Sie sollten über umfassende Kenntnisse in medizinischer Statistik und Grundkenntnisse in Humanmedizin und klinischer Pharmakologie verfügen. Ebenso besitzen Sie sehr gute, praktische Kenntnisse in der Datenanalyse mit Hilfe der Software SAS.

Von großem Vorteil sind Erfahrungen mit der Planung und Auswertung von Studien. Sie haben Freude an Projektarbeit in internationalen Teams sowie Interesse an klinischer Arzneimittelentwicklung und verfügen über ein hohes Maß an Einsatzbereitschaft und Belastbarkeit.

Wir bieten:

Leistungs- und anforderungsgerechte Bezahlung, die vielfältigen Sozialleistungen eines Großunternehmens sowie gute Chancen zur Weiterentwicklung sind für uns selbstverständlich. Es erwartet Sie ein hochmotiviertes Team mit internationaler Ausrichtung.

Haben wir Ihr Interesse geweckt?
Dann freuen wir uns auf Ihre Bewerbung

Hoechst Marion Roussel
Deutschland GmbH
Bewerberservice
Gebäude K 607
65926 Frankfurt

Hoechst

Hoechst Marion Roussel
Das Pharma-Unternehmen von Hoechst