Bioequivalence and Generic Medicines

by Dr. Pieter Zanen

Introduction

Nowadays bioequivalence studies are a pivotal part of registration dossiers. These studies measure the bioavailability of two (or more) formulations of the same active ingredient. The purpose of the study is that the bioavailability of the formulations under investigation is shown to be equal. Based on that conclusion, one may subsequently claim that the therapeutic quality of these formulations is identical. The latter means that both the beneficial and side effects are identical and hence the formulations are truly interchangeable.

Equivalence studies are necessary in two cases:

1. For the formulation and manufacture of generic medicines. Logically regulatory bodies require that the therapeutic quality elicited by the generic formulation is identical to that of the originator formulation. The patient must experience the same efficacy from the generic and the originator formulation: interchangeability.

2. For formulation changes by the originator. A good example of the latter is the replacement of CFC's in MDI's (metered dose inhalers) by the ozone-friendly HFA-propellants. Such major change of a formulation must incur no change in the therapeutic quality, which must be proven by the manufacturer. In other words, an equivalence study is obligatory.

The emergence of the generic industry in recent years means that currently most bioequivalence studies are carried out by this part of the pharmaceutical industry. Experience with these studies is vast within the generic industry and many guidelines or requirements are the result of the scientific discussion these companies have with regulatory bodies like the EMEA (European Medicines Evaluation Agency) and the FDA (US Food & Drug Administration). **However, one must realise that for the originator industry the same rules do apply, so no distinction can be made between the generic and originator industry in terms of how to perform a bioequivalence study. The standards for these studies are accepted by both the originator and generic industries.**

In this review we will explain the approach to bioequivalence. We start with the basis underlying the comparison of the plasmalevel-time profiles to claim therapeutic equivalence. After that we will outline the pharmacokinetic calculations and go into some depth on the statistical evaluation of the data. The latter is an essential part of any bioequivalence study.

Using plasmaconcentration-time profiles to claim therapeutic equivalence

When a major formulation change occurs or a generic equivalent of an originator formulation desires market approval, a clinical comparative study in patients is the most obvious route to show therapeutic equivalence. In such a study, one simply compares the pharmacodynamic or the therapeutic effects of the two formulations (and/or other relevant clinical endpoint[1] like side effects). When both beneficial effects and side effects are the same, the formulations are interchangeable.

Frequently this proves to be a very (or even too) difficult task. Clinical comparative studies are often hampered by a lack of clearly defined and measurable endpoints. Take antidepressants for example: the measurement of the severity of a depression is a science in itself and a consensus on the best measurement method often does not exist.

A second problem is that clinical studies often show a huge variability of the measured endpoint. Without going into too much detail now, when one calculates the sample size for trials suffering from high variability, it means huge numbers of volunteers or patients. Many trials will therefore not be sufficiently powerful to detect differences, which endangers the concept of no therapeutic change. It is quite easy to design a trial in which no differences can be shown, when variability is high. It is also easy to conceive that a trial, which cannot detect small differences between two formulations, leaves a large difference undetected. Such major, but undetected difference could be clinically very relevant.

The conclusion is that clinical studies are not the optimum choice for comparisons of formulations (where small differences are to be expected) and an alternative method had to be developed, which is the pharmacokinetic approach. The advantages of a pharmacokinetic approach are the superb definition of the endpoint (the plasmaconcentration of the drug) and the lower variability of that endpoint. These characteristics solve many of the problems encountered with clinical testing: they lead to smaller and more powerful trials, which are to the benefit of both the manufacturer and the patient.

The drug receptor interaction

The basic assumption underlying the kinetic approach of bioequivalence studies is that when the same number of drug molecules occupy the same number of receptors, identical pharmacodynamic effects will be elicited. So the problem of showing identical clinical effects is now replaced by showing that identical numbers of drug molecules are present on the receptors at any time point, irrespective the formulation used to deliver the molecules. Now within the same subject the numbers of receptors is of course stable, so any difference in therapeutic effect is caused by differences in the number of drug molecules delivered to the receptor.

The advantages are clear: the measurement of the clinical (or side effects) is of no interest anymore. It is even of less interest what exactly the clinical effect(s) is or are. All these effects are due to a drug-receptor interaction: multiple effects might be due to one type of interaction or more than one type of receptor might elicit more than one effect. As long as one can prove that the number of drug molecules present at receptor X at time point Y are identical after administration of formulation A or B, identical clinical and side effects must exist.

[1] An endpoint in a study is the parameter one chooses to measure.

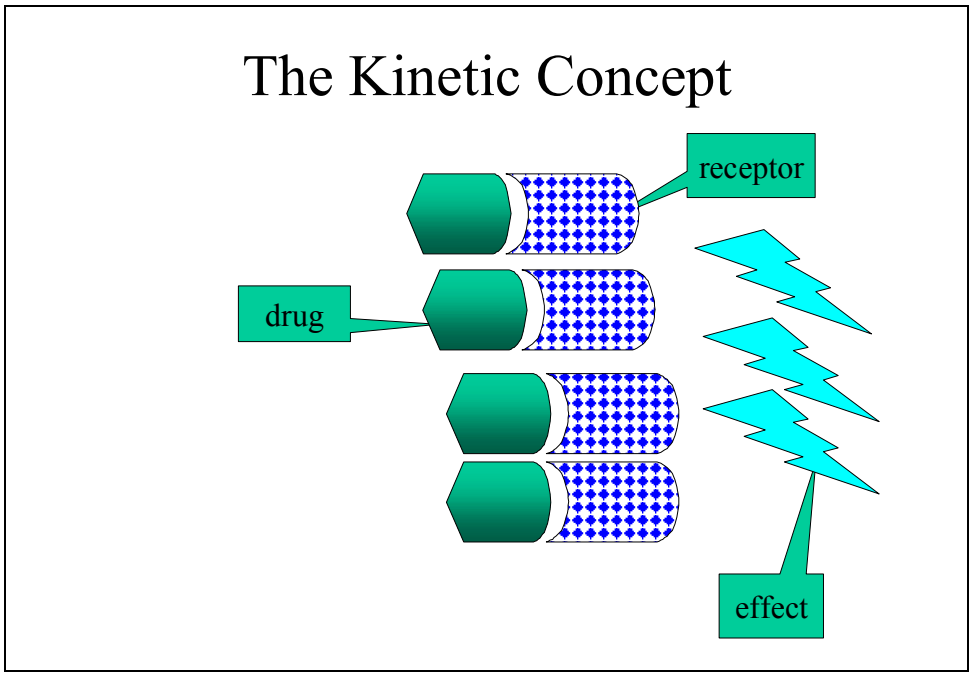

Fig. 1 The effect of a drug is related to the number of receptor occupied by drug molecules

Having defined bio-equivalence as proving that the same number of drug molecules occupy the receptors, one can further simplify the problem by looking at the parameters which influence the number of drug molecules at the receptor.
The systemic circulation delivers the drug molecules to the receptor, so the number of molecules in the systemic circulation is a measure for the number at the receptor. The parameters governing the plasmaconcentration of a drug are absorption, distribution, metabolism and elimination of the active drug. So when these parameters are measured and shown to be equal, the numbers of drug molecules which reach the receptors are equal. As a result one can only conclude that the therapeutic effects must be equal too.

Basic to this approach is that the human body handles a specific drug in a specific way, i.e. absorption, distribution, metabolism and elimination of a drug molecule is of course always the same, irrespective the source of the drug. Whether it is administered using formulation A or B, does not matter: the chemical nature of the drug determines the pharmacokinetic handling by the body. The latter may show inter-individual differences but these are of no interest at all, because the basic goal of bioequivalence testing is to show interchangeability in the individual patient (only intra-individual variability counts).

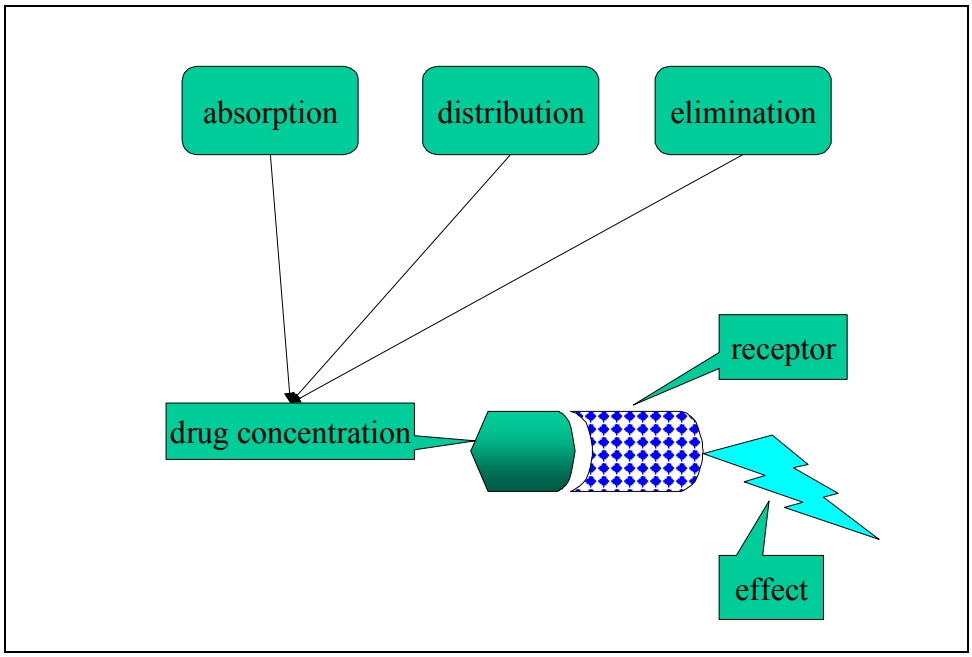

Fig. 2 The number of drug molecules present at the receptor is governed by kinetic processes like absorption, distribution, metabolism and elimination

However, distribution, metabolism and elimination cannot be influenced by formulation A or B. These parameters are totally dependent on the nature of the drug or on the genetics of the patient. So when the chemical nature of the drug in two different (generic and originator) formulations is identical, the distribution and elimination patterns are exactly the same. So any change in the number of drug molecules at the receptor can only be caused by differences in absorption from formulation A or B.

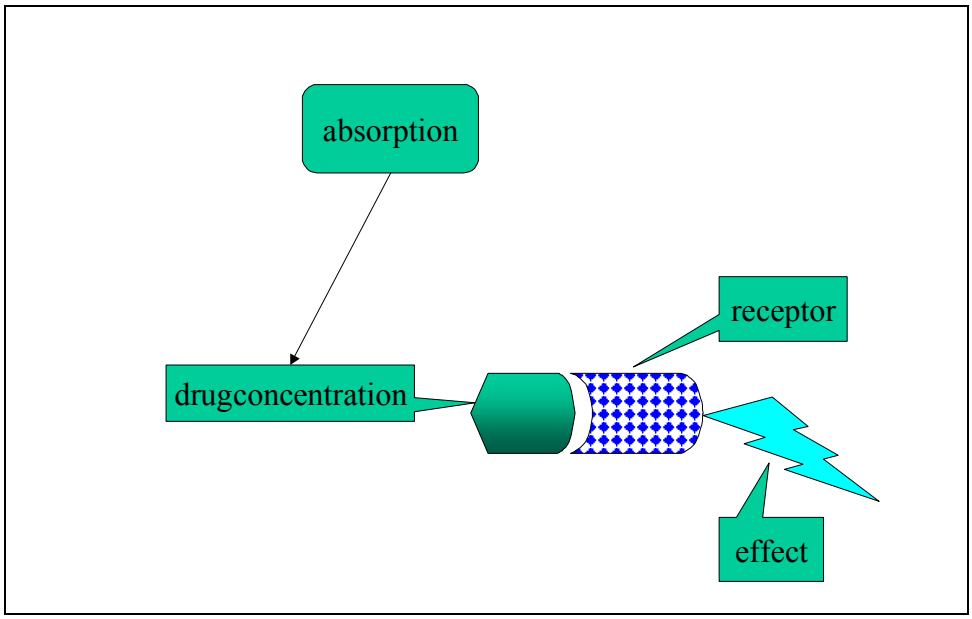

Fig. 3 In all patients distribution, metabolism and elimination are drug specific processes and not formulation dependent. As such they are constant and can be ignored in a comparative trial

It is now a fair assumption that the nature of the absorption process is also chemically determined, one must acknowledge that the gastro-intestinal tract will absorb a generic or originator drug molecule in exactly the same way. So if a difference in absorption does exist, this can only be caused by changes in delivery of the drug from the formulations:

disintegration of the dosage form and dissolution of the drug. The latter two are known under the term ‘pharmaceutical availability’.

The pharmaceutical availability is under the direct control of the manufacturer. In other words, the kinetic approach to bioequivalence studies can be rephrased as follows: absorption, distribution, metabolism and elimination are constant within the same volunteer, so differences in the plasmaconcentrations are due to differences in the pharmaceutical phase. A bioequivalence study is simply a check on identical pharmaceutical phases. Variability of the absorption, distribution, metabolism and elimination are therefore disturbance factors increasing the total variability of plasmaconcentrations thereby leading to an increase of the sample size.

To summarise: the plasmaconcentration of a drug determines the number of drug molecules at the receptor and hence the therapeutic effect. The plasmaconcentration is governed by absorption, distribution, metabolism and elimination. The latter are constant within the same subject, so differences in the plasmaconcentration (= therapeutic effect) are due to differences in the amount of drug absorbed which, in turn, depends on the delivery of drug from the formulation.

The parameters measured in a bioequivalence study

Earlier it was argued that a bioequivalence study is a check on the similarity of the release characteristics of formulation A and B. The amount of drug molecules released and the speed of the release are therefore the most important parameters. Rephrased: the rate and extent of the release. In the in-vivo bioequivalence study these characteristics are determined by measuring the following parameters:

1. the area under the plasmaconcentration-time curve (AUC), because it describes the total number of molecules present in plasma, thereby informing the researchers on the extent of the release;

2. the maximum plasmaconcentration (C_{max}) because it is linked to the speed of the release;

3. the time at which the maximum plasmaconcentration is reached (T_{max}) because it is linked to the speed of the release;

4. the elimination half life $T_{1/2}$, because it is linked to the elimination of the drug (see later). $T_{1/2}$ is obtained by calculation of K_{el}, the elimination constant.

The last three factors fully determine the shape of the plasmaconcentration-time curve and strategies to compare the shape of the curve itself instead of these 'derived' parameters are of little use.

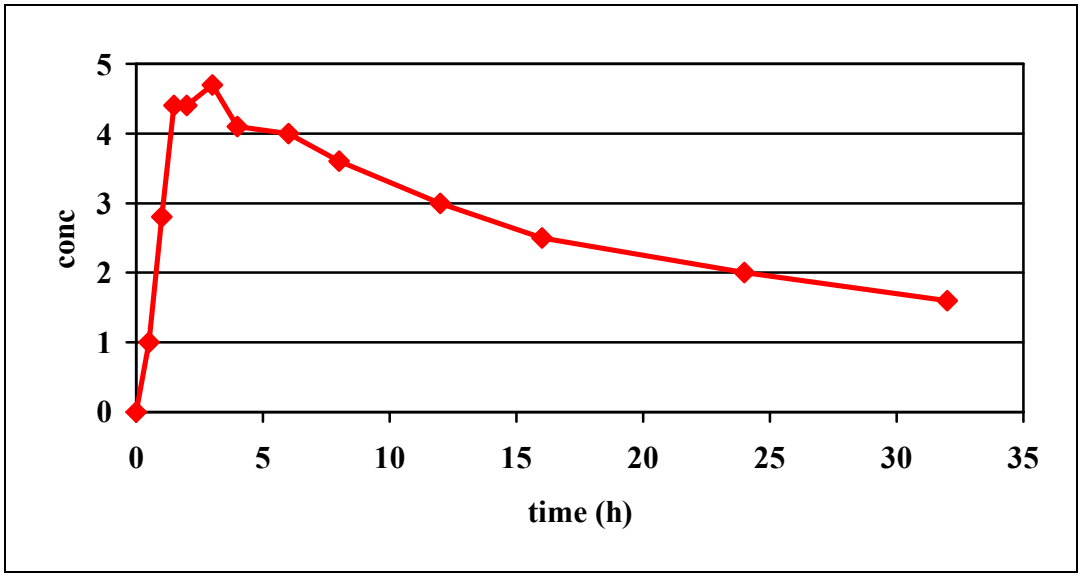

Fig. 4 Typical plasmaconcentration-time curve showing time of the X-axis and the plasmaconcentration on the Y-axis. The highest concentration is Cmax

Fig. 4 is a typical example of a plasmaconcentration-time profile of a drug in a volunteer. One can distinguish an absorption and elimination phase. When the mass absorbed equals the mass eliminated, C_{max} is present. Before C_{max} is reached (before T_{max}) the absorption is higher than the elimination, after T_{max} the situation is reversed.

Frequently elimination is a so-called first order process, which means that per unit of time a percentage of the mass present in the blood disappears from it. So for example every hours 5% of the mass present in the blood disappears, which means that as the plasmaconcentration declines, the eliminated mass per time unit also declines.

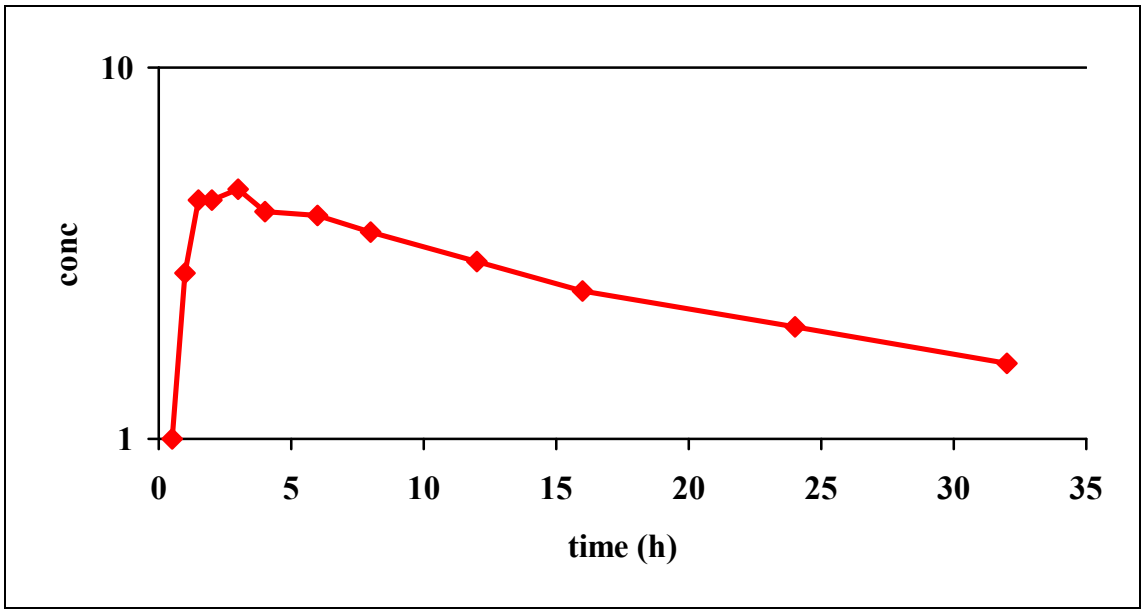

Fig. 5 An ln-transformed plasmaconcentration-time curve, showing a 'linear' elimination phase (actually the elimination is an exponential function)

When elimination is a true first order process, a log transformation of the measured plasmaconcentrations will render a straight line during the elimination phase, which we shall use later to calculate the elimination half-life.

Calculation of the AUC_{0-t}

$$AUC_{0-t} = \sum_{i=1}^{t} \left(\frac{C_i + C_{i-1}}{2} \right) (t_i - t_{i-1})$$

Eqn 1 Calculation of the area under the plasmaconcentration-time curve.

The AUC is calculated by taking the average of two subsequent plasmaconcentrations (C_i *and* C_{i-1}) and multiplying that average by the time difference between the consecutive measuring points (t_i and t_{i-1}). All these outcomes are then summed to render the AUC from 0 to the last measuring point. This approach is called the *linear trapezoidal* approach.

The measurement schedule (= sampling schedule) must be designed in such a way that the absorption of the drug is adequately charted, so minimum requirements for the length of the sampling scheme exists.

Note that the contribution of widely separated measuring points to the total AUC can be strong, because t_i minus t_{i-1} is then quite large. Measurement errors made during this phase have a significant influence on the results of the study. Figure one shows that the plasmaconcentrations in the elimination phase are slightly higher than they should be due to measurement errors. The AUC is increased considerably in a incorrect way: to overcome this hurdle the log-trapezoidal rule is used.

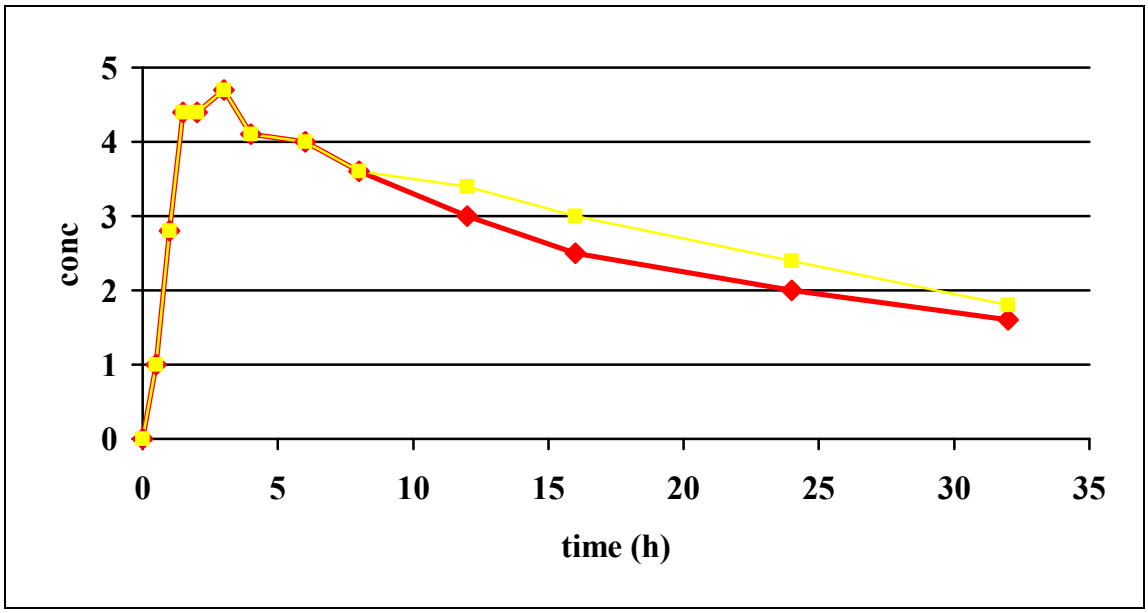

Fig. 6 Comparison of two plasmaconcentration-time curves. The yellow upper curve shows higher concentrations at later time-points due to a measurement error, this increases the AUC considerably.

$$AUC_{0-t} = \sum_{i=1}^{t} \left(\frac{C_i - C_{i-1}}{\frac{1}{\Delta t} \ln\left(C_i / C_{i-1}\right)} \right)$$

Eqn 2 Calculation of the area under the plasmaconcentration-time curve, using the log-trapezoidal rule.

Calculation of C_{max}

The calculation of C_{max} is, luckily, very simple: select the highest value.

Calculation of T_{max}

The calculation of T_{max} is equally very simple: select the time point at which C_{max} occurs.

Calculation of K_{el} and $T_{1/2}$

The calculation of K_{el} is an essential part of any bioequivalence study. Above the AUC_{0-t} was calculated but there is a requirement that the absorption phase of the drug has to be adequately described. Generally this is the case when the AUC_{0-t} is >0.8 of the extrapolated $AUC_{0-\infty}$. The latter parameter cannot be measured, of course, but is estimated and for that estimate the K_{el} is needed. When AUC_{0-t} is >0.8 of $AUC_{0-\infty}$ it simply means that the sampling scheme was sufficiently long to be sure that the absorption phase of the drug is indeed adequately described. (For drugs with a very long half-life this requirement might be relaxed). As we discussed earlier, elimination is a first order process and a natural log (ln)-transformation makes it possible to draw a straight line through the elimination phase. The slope of the regression line is now equivalent to K_{el} or the elimination constant.

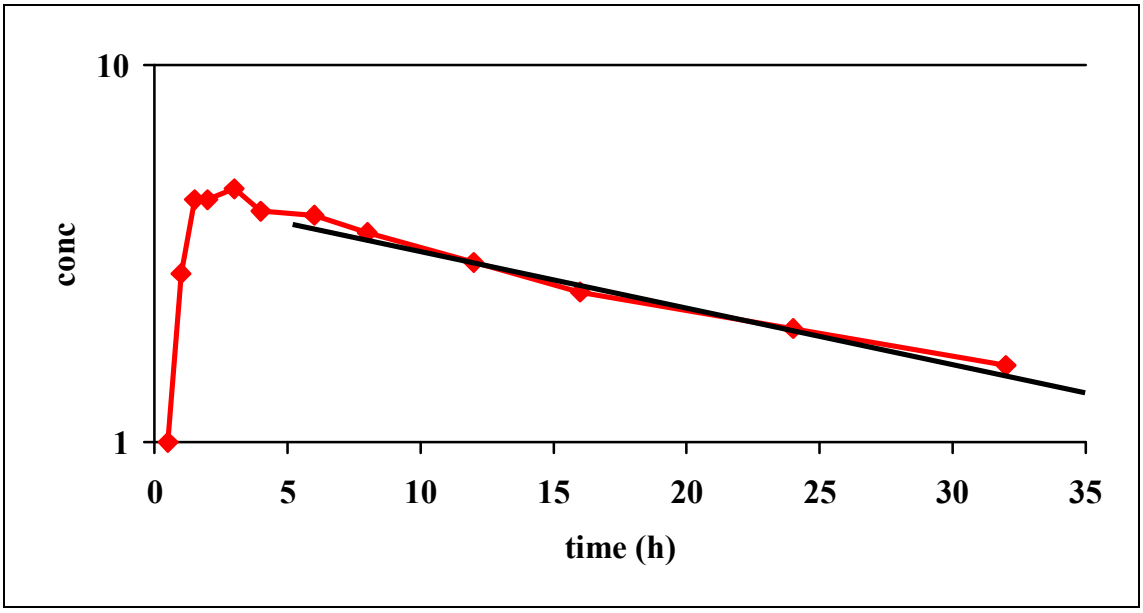

Fig. 7 Ln-transformed plasmaconcentration-time curve showing a regression line through the elimination phase. The slope of the regression line represents the elimination constant.

The calculation of the half-life is now rather simple. One simply divides 0.693 by the K_{el} to obtain the $T_{1/2}$. The relationship between K_{el} and $T_{1/2}$: $T_{1/2}= 0.5/k_{el}$. The term 0.693 is derived from ln(0.5)=0.693 (ignoring signs): K_{el} describes the lowering of the ln-transformed plasmaconcentration per time unit and so we have to ln-transform 0.5 too to obtain the correct estimate of $T_{1/2}$.

$$T_{1/2} = \frac{0.693}{K_{el}}$$

Eqn 3 Calculation of $T_{1/2}$

Calculation of the $AUC_{0-\infty}$

The next step in process is to extend the plasmaconcentration-time profile to infinity to obtain the $AUC_{0-\infty}$. The latter parameter is an estimate of the total mass of drug present in the blood and also serves as a guide for adequate sampling.

To do so, the K_{el} is of course the most logical parameter, next to the last plasmaconcentration. As mentioned before, K_{el} describes the loss of drug per time unit (/h) so division by of C_{last} (mg/l) results in a measure with the unit mg/l*h, which is the unit for an AUC. The outcome of this calculation is the AUC from t_{last} to infinity ($AUC_{t-\infty}$), so to obtain the $AUC_{0-\infty}$ one has to add AUC_{0-t} and $AUC_{t-\infty}$

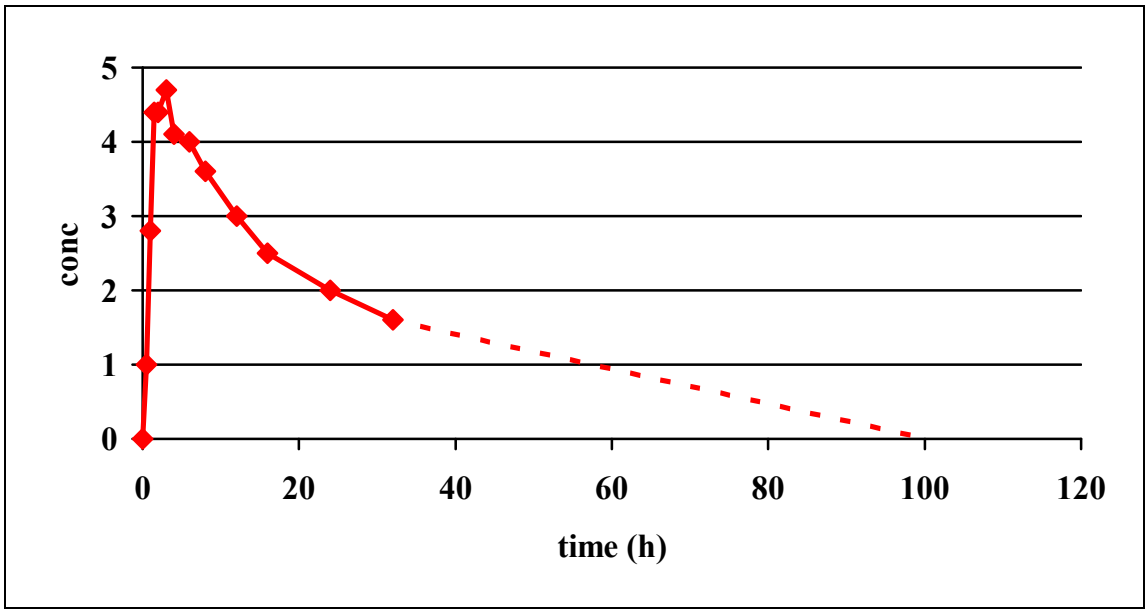

Fig. 8 Estimation of the $AUC_{0-\infty}$

$$AUC_{0-\infty} = AUC_{0-t} + \frac{C_{last}}{K_{el}}$$

Eqn 4 Calculation of the $AUC_{0-\infty}$

Some alternative parameters in use

Normalised C_{max}

The C_{max}, T_{max} and AUC are the classical parameters in use to establish bioequivalence. Parameters, like C_{max} and T_{max} suffer however from considerable variability, rendering them less effective. The normalised C_{max} is used by some: it is simply the C_{max} divided by the AUC. Research has shown that the normalised C_{max} is less dependent on intra-subject differences and on differences in the extent of absorption. The latter is of course logical because in a case of lower absorption both the C_{max} and the AUC will be affected to the same degree. Their ratio therefore does not change. The normalised C_{max} also turns out to be more powerful when true equivalence does exist, but also detects differences more quickly.

Mean residence time

The mean residence time (MRT) is a good indication for the time a drug molecule is present in the body without being metabolised or eliminated. In case of so-called sustained release preparations the MRT is often used. Its calculation is, however, tedious. It requires first an estimate of the time the molecule spends in the human body, which is obtained by calculation of the so-called area under the moment curve (AUMC), the latter is equal to the area under the curve (AUC) but includes the timing information. The equation for the $AUMC_{0-t}$ is: $AUMC_{0-t} = \sum_{i=1}^{t}\left(\frac{t_i C_i + t_{i-1} C_{i-1}}{2}\right)\left(t_i - t_{i-1}\right)$. The AUMC is determined by multiplying the plasmaconcentration with the time value before adding them together and dividing by 2. In other words the calculation is very similar to that of the AUC, the only difference being the

multiplication of the plasmaconcentration by time. The equation for the terminal part of the AUMC is: $AUMC_{t-\infty} = \frac{t_{last} \times C_{last}}{K_{el}} + \frac{C_{last}}{K_{el}^{2}}$. It uses the last timepoints and the plasmaconcentrations as well as K_{el} to arrive at the correct value. Subsequently as expected one can add the $AUMC_{0\text{-}t}$ to obtain the $AUMC_{0\text{-}\infty}$. Division of the $AUMC_{0\text{-}\infty}$ by $AUC_{0\text{-}\infty}$ results in a time value which is equivalent to the mean residence time of a drug the body: $MRT = \frac{AUMC_{0-\infty}}{AUC_{0-\infty}}$.

Parameters in use with sustained release preparations

Bioequivalence studies are also obligatory for sustained release preparations, which emit the drug over longer periods of time. The testing of these formulations is more difficult because the most important goal of these formulations is to obtain smooth steady state plasmaconcentration-time profile. It is therefore not surprising that the mean goal for sustained release preparations bioequivalence studies to show similarity of the steady state plasmaconcentration-time profile.

The classical single dose parameters maintain their validity, but no longer describe the plasmaconcentration-time profile adequately. New added parameters are the C_{min} and the % peak trough fluctuation. C_{min} describes the lowest plasmaconcentration just before the next dose is given and the (percent) peak trough fluctuation the swing of the plasmaconcentration-time profile between two administrations. The latter is calculated by $\%\mathrm{PTF} = \frac{(C_{max} - C_{min})}{C_{avereage}}$. The denominator is $C_{average,}$ which is the average plasmaconcentration over the dosing interval. Division by C_{min} has the problem that the parameter becomes unstable, especially when C_{min} is small. Tiny changes in a small C_{min} will have large effects on the calculated value. The reverse will be true for division by C_{max}, so division by $C_{average}$ seems the wisest thing to do. $C_{average}$ is calculated by $C_{aver} = AUC_{0-t} / t$, where t= the dosing interval.

The general approach to bioequivalence testing

Equivalence limits

Equivalence testing has some very specific problems and approaches, which renders the statistical evaluation rather typical. The major characteristic of equivalence testing is that not every difference between two formulations (A and B) automatically leads to the rejection of an equivalence claim. This is best illustrated by an imaginary study in which two originator formulations from the same batch are tested: one we call 'reference' and the other 'test'. In this study equivalence is present by definition. Still it is very well possible that the mean AUC's of the reference and test formulation will differ somewhat. This is of course due to random error, measurements errors and other disturbing influences.

These differences are acceptable and do not hinder a claim of equivalence between the two formulations. This line of reasoning illustrates that a difference between two formulations is composed of several components. One component being mere random error, which is unavoidable and hence acceptable. Systematic differences between formulations are another possible component, but point to a different behaviour of the reference and test formulation. This component is not acceptable, although it can be agreed that small differences are of no clinical importance.

The typical *a priori* problem of equivalence studies is therefore that one has to define a difference between two formulations which will not cause therapeutic inequivalence. This maximal allowable difference or the equivalence limit is a pivotal choice: when one chooses a very large or wide limit, the actual difference between formulations can be that large, that it is clinically relevant. The medical profession will then judge that true therapeutic equivalence does not exist: the so-called consumer risk for inequivalence is high. Choosing a wide limit is however advantageous in terms of sample size: equivalence studies with wide limits are small and inexpensive trials and hence appealing to many (see later sections).

On the other hand choosing very narrow limits may be appealing because then one is very certain that equivalence does exist. However, rejecting equivalence because small differences between formulations do exist is not scientifically sound when one agrees that these narrow limits are clinically uninteresting.

The choice of the equivalence limits is a purely clinical debate and not a statistical one. Medical doctors have to decide what is an acceptable difference between formulations, which does not lead to therapeutic relevant differences. The statistician can be of help in this choice process and will calculate the sample size, but can do so only with the input of medical doctors.

Reducing consumer risk

The first equivalence studies followed a now-abandoned approach called the *power approach*. It is interesting to outline this approach because it gives insight into the current one, which again is typical for equivalence testing. The power approach recognised that the difference between two formulations may not exceed a predefined and clinically significant difference (=equivalence limit). So the trial had to be aimed at finding that predefined difference. When this clinically relevant difference is not present, the true differences between the two formulations must be smaller and hence equivalence must exist. *A significant difference between formulation hence means inequivalence.*

There are some pitfalls here. It is very easy to show a lack of significant differences: a trial with a low statistical power is able to pick up only (very) large differences. It was therefore

stated that the trial should have sufficient power, that is at least 80% power to pick up the predefined difference between formulations. Following this route of the power approach, the calculation of the sample size is straightforward. The well-known equation for the sample size (in a cross-over situation) is used.

$$N = 2\left(Z_{\alpha/2} + Z_{\beta} \right)^2 \frac{\sigma_d^2}{\delta^2}$$

Eqn 5 Sample size equation for a cross-over study. σ is the standard deviation of the difference δ

The delta-value in this equation equals the predefined maximal allowable difference between the formulations, while sigma represents the standard deviation of the difference (in a cross-over study). Again, the key element of the power approach is that, when no significant differences are shown in a sample of a certain size, the true difference between the formulations is smaller than the maximally acceptable one (=equivalence limit) and so equivalence must exist.

The reasoning seems sound, but the power approach lacks sufficient consumer protection: the risk of a false equivalence claim is high.
The basis of the method is a sample size calculation, which should guarantee sufficient discriminative power to detect the predefined difference. The problem with any sample size calculation is, that it is always based on data previously obtained. Better said, based on data from a sample of subjects, which by definition is not the sample one uses in a next study. It is conceivable that variability in the actual sample is larger compared to the earlier sample variability and as result the current trial will lose discriminative power. The latter is equivalent with a reduced probability to detect significant differences. It is hard, in daily practice, to select a new sample which is an exact copy of the previous one.

It is also possible (in theory) for a manufacturer to influence *a priori* the outcome of its study by selecting volunteers (or reference formulations) with a higher variability than was present in the previous study. This wicked scheme reduces the probability to detect significant differences, reduces patient protection because it is possible that non-equivalent formulations will be approved.

An obvious solution is to calculate the power of the study under consideration and see whether it is indeed 80%. This, however, is impossible. At the moment that the two mean AUC's of the formulations are very similar or close to each other, the power of the trial is by definition low or even zero. The equation in use to calculate the power of a study is $\sqrt{\frac{N}{2}} \times \frac{\delta_d}{\sigma_d} \quad Z_{\alpha/2} = Z_{\beta}$ and it is easy to see when δ is small the power of the study is by definition low. So the 80% power to detect a relevant difference is purely based on history and frequently there is no strong evidence that the current sample is identical to one used as reference which was used for sample size calculations.

A solution to the above problem is to include a large number of volunteers. At first glance a logical step, but the drawback is that an increase of the sample size means in increase of discriminative power. As a direct consequence there is an increased probability of finding significant differences where one should not. The manufacturer is punished for his honesty by the task to explain that the significant difference found is not a clinically relevant difference, but due to too large a sample size.

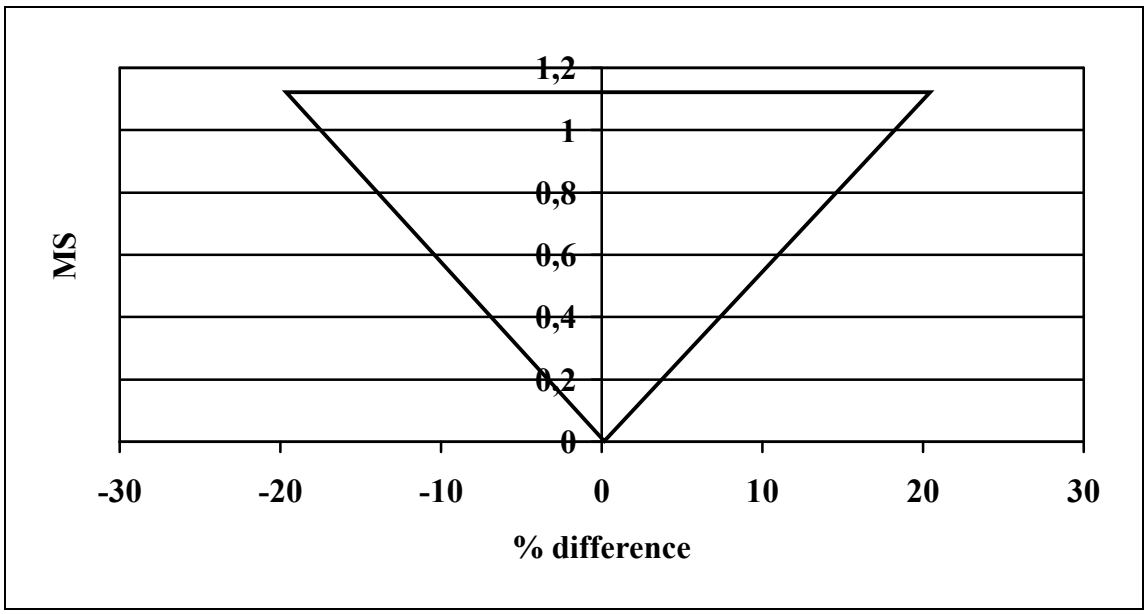

Fig. 9 Acceptance area as a function of the percentual difference between two formulations and the standard deviation, here depicted as the MS, in the power approach. The inverted triangle 'causes' small differences between formulations to be significant when the standard deviation is small (=inequivalent). On the other hand large differnces will not cause significant differences when the standard deviation is large (=equivalent).

The situation outlined above is illustrated in the above graph. On the X-axis the difference between two formulations is depicted as a percentage. On the Y-axis the standard deviation (=error mean square of the analysis of variance). The triangle depicts the region of equivalence: as long as one stays within that region equivalence can be claimed.

Let us start at the bottom or tip of the triangle. Here according to the power approach any small difference between the formulations will quickly render significant differences between them, because the error MS is so small. This is generally the case when the sample size is high.

At increasing error MS, the differences between the formulations may increase without a conclusion of inequivalence because it becomes harder and harder to detect differences. In other words the difference between formulations (as a percentage) may increase, so the region of equivalence becomes larger and larger. This ends up in a triangle shape. (Shuirmann from the FDA showed that there is an upper limit to the increase of the error MS, which is depicted by the base of the triangle. The error MS is then so large that it is impossible to reach 80% power).

Schuirmann argued that a reversal of the triangle is a much better situation. That new approach is illustrated in the graph below. On the X-axis again the difference between two formulations is depicted as a percentage and on the Y-axis again the error MS. The triangle again depicts the region of equivalence: as long as one stays within that region equivalence can be claimed. (For the insider, the basis of the triangle is the equivalence limit.)

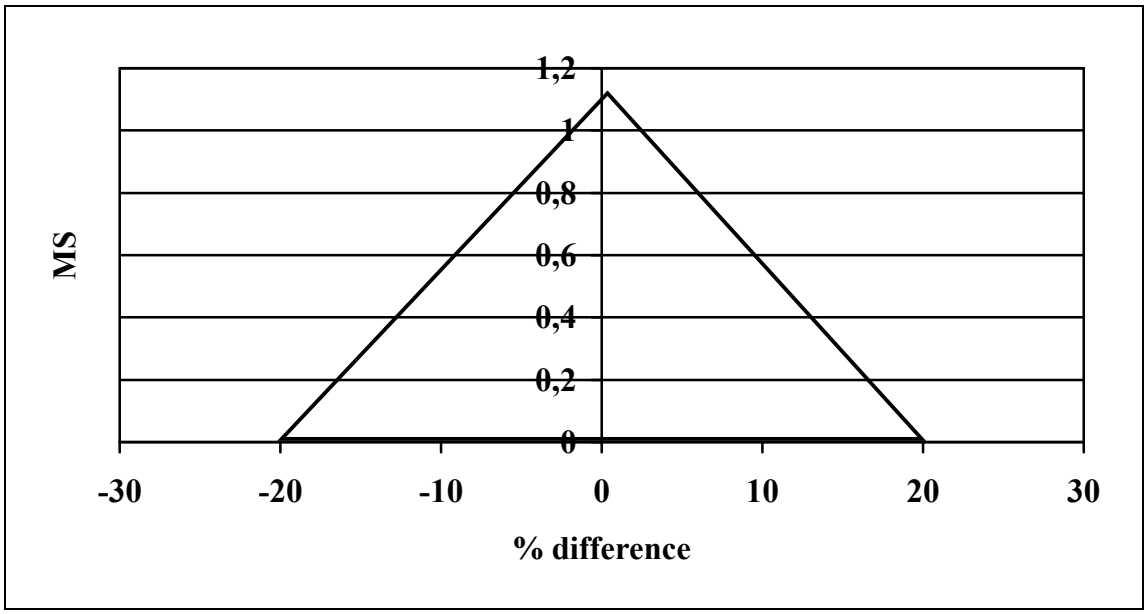

Fig. 10 Acceptance area as a function of the percentual difference between two formulations and the standard deviation in the Schuirmann approach. The triangle now leads to a starightforward equivalence claim when the standard deviation is small. On the other hand large differences will not lead to such claims.

The big improvement is that when variability is high (large error MS), the probability of claiming equivalence will reduce enormously, thereby solving the major disadvantage of the power approach sketched above. Let us start at the bottom of the triangle. Here again there is hardly any difference between the formulations (0%) and the Schuirmann approach will render equivalence between the formulations even when the error MS is small. It is also easy to see that equivalence can still be claimed when the difference between formulations is not 0, but smaller than the equivalence limit (here ±20%).
At an increasing error MS (= smaller sample size or huge variability), the differences between the formulations must be small because it will become harder to claim equivalence. This ends up in an inverted triangle.

As one can see all disadvantages of the power approach are solved now: a huge variability makes it very hard to claim equivalence and a small variability makes it easy. The manufacturer is rewarded for setting up an adequately sized and well-controlled trial: it more or less guarantees an equivalence claim and hence approval by the regulatory bodies. The patient protection against false claims is also increased because studies inadequately controlled or sized will not render an equivalence claim and regulatory approval. Such a product will not enter the market.

— H_{01}: $\mu_{test}-\mu_{ref} \leq -0.20\mu_{ref}$

— H_{02}: $\mu_{test}-\mu_{ref} \geq 0.20\mu_{ref}$

— H_1: $-0.20\mu_{ref} < \mu_{test}-\mu_{ref} < 0.20\mu_{ref}$

Eqn 6 Null and alternative hypothesis under the Schuirmann approach ($0.2\mu_{ref}$ is the equivalence limit).

The principles laid down by Schuirmann are translated in the following approach which states that the difference between reference and test formulations may not exceed certain limits, the so-called equivalence limits (here formulated as 0.2x μ_{ref}) . If that is the case the null hypotheses, which states that the differences are larger than the equivalence limits, is not rejected. The alternative hypothesis now states that the difference is smaller than the equivalence limits dictate and hence equivalence can be claimed.

The advantage of this approach is that the null hypothesis is only rejected in favour of the alternative hypothesis when the power of the trial is sufficiently high. The error MS or standard deviation must be low or the sample size sufficiently high. It is a characteristic of this method that when two means are equal, the null hypotheses are quickly rejected using a low number of volunteers.

Below we graphically depict that the Schuirman approach states that the difference between reference and test may not exceed certain limits, the so-called equivalence limits.

If μ_{test} is larger than μ_{ref} the difference has a positive value, it may not exceed θ_{upper}, when the difference has a negative value it may not exceed θ_{lower}. In the typical way of speaking: one accepts the null hypothesis or inequivalence when μ_{test} minus μ_{ref} is larger than θ_{upper}, but when μ_{test} minus μ_{ref} is smaller than θ_{lower}, we reject the null hypothesis or inequivalence and accept the alternative hypothesis or equivalence.

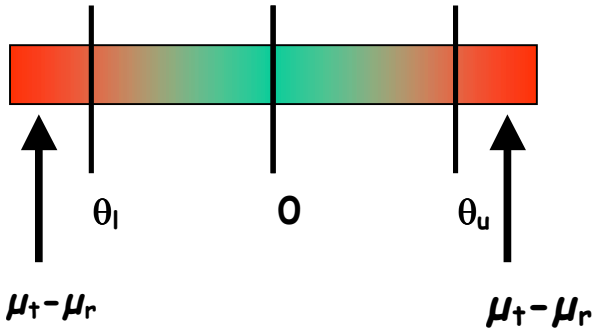

Fig. 11 Graphical presentation of the Schuirmann approach: when the difference between the mean AUC of two formulations is (or can be) larger than a predefined allowable difference, equivalence can not be claimed.

We now restate the approach by bringing the μ_{ref} term to the other side: it then reads as 1] one accepts the null hypothesis of a negative difference between the two formulations plus the value of the equivalence limit (in this case 0.2 x μ_{ref}) is smaller than or equal to zero and 2] one accepts the null hypothesis of a positive difference between the two formulations minus the value of the equivalence limit (in this case 0.2 x μ_{ref}) is greater than or equal to zero. Again rejection of both hypothesis renders equivalence.

The decision whether to accept or reject the null hypothesis is based on the calculation of the T-value. Remember that one must reject the null hypothesis, so there must be a significant T-value present. One states that **inequivalence** is present when 1] the T-value of the first (negative) term does not exceed the T-value of significance or when 2] the T-value of the first (positive) term does not exceed the T-value of significance.
The calculation of the T-value is as follows:

$$\left| \frac{\mu_{test} - \mu_{ref} \pm 0.2\mu_{ref}}{SE} \right| \leq t_{\alpha,df}$$

Eqn 7 Calculation of the T-value of the difference between the AUC's of two formulations ana when that T-value is smaller than the critical value H_0 is not rejected and inequivalence must be concluded (α=0.05, df is degrees of freedom of the error MS).

Where $t_{a,df}$ is the critical T-value from the Student T-distribution and SE is the standard error (derived from the error mean square from the analysis of variance).

There is a popular variant to the above equation. It is easy to see that a small rearrangement of the terms renders a new equation and the upper term becomes the equation for a confidence interval. When the term μ_{ref} is again brought to the other side to obtain 0.2. μ_{ref}, it can be stated that the confidence interval must lay within the equivalence limits (which now can be called the equivalence interval).

$$\frac{\left(\mu_{test} - \mu_{ref}\right) \pm t_{\alpha,df}.SE}{\mu_{ref}} \leq 0.2$$

Eqn 8 Reformulated approach to equation 7 (α=0.05, df is degrees of freedom of the error MS)

In daily practice many multiply the above terms by 100% to obtain a percentage expression. It then reads as that the confidence interval must lay within a range ±20% of μ_{ref}. When the latter is indexed as 1 (or 100%), the equivalence range runs from 0.8 to 1,2 (or from 80% to 120%).

It is recognised that it is not possible for a test AUC to be larger of smaller at the same time. So one-sided tests will do. Under 'normal' circumstances in statistical testing one departs from an assumption that the differences between formulations can be positive or negative. In other words that the test mean can lay at either end of the Student T-distribution. With regard to the AUC's it is not possible that one time the difference between test and reference is negative and the next time it is positive: it is not a random variable. It always has the same direction, so one-sided tests are sufficient in this case. The tests are done at an α-level of 5% and so a one-sided 95% confidence interval is obtained. Those who still test two-sided therefore obtain a 90% confidence interval, but this is not necessary, of course.

To summarise the current approach basically calculates the 95% confidence interval of the difference in the measured parameters like AUC, C_{max}, etc. As known the 95% confidence interval is the range of values in which, with 95% probability, the true or population difference lies. That interval must fit within the equivalence limits, which means that with 95% probability equivalence does exists: the consumer risk of inequivalence is therefore limited to 5%. When the manufacturer has chosen a large sample size, the 95% confidence interval of the differences will shrink and hence the consumer risk of inequivalence is reduced below 5%.

What is also very important is that a large difference between the formulations automatically leads to a need for a (very) small confidence interval. An example to illustrate: suppose that the equivalence limit dictates a maximum difference of ±5. The actual difference between the formulations is 0, so the width of the 90% confidence interval may be ±5 (it will still fit within

the ±5 range). A next formulation shows however a difference of +2, so now the confidence interval may only be ±3 in order to fit. The smaller interval can be reached by increasing the sample size: the larger the difference is between two formulations the more accurate the study must be in order to be very sure that the measured difference is indeed the true difference.

Study design, evaluation and sample size

Study design

Basically two types of designs are possible, that is the parallel and cross-over design.
The major difference between these designs is the way they deal with intersubject variability. Intersubject variability is a measure of the differences between subjects. On the other hand intrasubject variability is a measure of the differences within subjects. Both types of variability are present in each trial, but in the cross-over design the intersubject variability is eliminated. The subject functions as his or hers own control and a difference between formulations *within one person* is only influenced by the (non)random within variability. This makes the cross-over design much more efficient in terms of sample size. One should remember than sometimes the intrasubject variability is very high and in these cases the advantage of a cross-over design rapidly fades away. This happens with so called highly variable drugs, where a considerable intrasubject variability is present. Nifedipine and acyclovir are examples of these highly variable drugs.

The most common design for a cross-over trial is the well known AB/BA trial. It tells that one splits the entire sample of subjects randomly into two groups. Group 1 will receive the drug or formulations (frequently called the *treatments*) in the order A-B or reference-test and the other group in the order B-A or test-reference. These two orders are called the *sequences*, so any two-formulation trial is a two-sequence trial.

At the same time any two-formulation trial is also a two-*period* trial. In the first period 50% of the volunteers receive A or reference and 50% B or test. In the second period the order is reversed of course.

The italic items listed above are to be considered in the statistical evaluation of the cross-over trial. The periods and the sequences are not supposed to exert an influence on the measured parameters like the AUC, $T_{1/2}$ or any other one. When a significant period or sequence effect is noted, the study can be invalid. Treatments or formulations might be different, but are frequently not in bioequivalence studies. The last factor to be included in the evaluation is the subjects: if that is not done the inter-individual variability is not accounted for.

One of the problems with any AB/BA trial is carry-over. Carry-over is present when the effects of the drugs in period 1 are still noticeable in period 2. For a bioequivalence study this would be the case if the first plasmalevel before administration of the drug in the second period is not 0. If that is the case the washout between the two periods was not sufficiently long. Carry-over in a bioequivalence study is tested by evaluating the sequence effect, but there are problems related to carry-over testing.

The major problem with carry-over tests is that it is not a test for carry-over: it is a test for unequal carry-over. Unequal meaning that the sequence AB is affected in a different way than the sequence BA. If the carry-over is equal in both sequences, it cannot be detected. Also the carry-over test is a test comparing two groups and as such it is an unpaired T-test, which is less powerful. Many therefore do not test for carry-over because of the many problems related with it. The problem of carry-over is best avoided: a proper design of the cross-over trial is the best guarantee for not being bothered by carry-over.

In the evaluation of a bioequivalence trial the FDA and EMEA still favour the evaluation of the sequence term. One of the reasons the test is still run is that carry-over is not the only problem. In any cross-over trial a so-called treatment by period interaction can be present. This means that the effects of any treatment is influenced by the period. Let's use as

example the testing of two sunblockers in the first period (winter) and in the second period (summer). Both blockers will be more effective in the second period, which is not due to carry over, but due to a period by treatment interaction. In a bioequivalence trial such interaction may be present in the absence of any carry-over. The problem is that one cannot discriminate between carry-over or a period by treatment interaction. These two can never be separated from each other.

Randomisation

A proper design is characterised by a complete randomisation of the subjects over the treatment arms. In bioequivalence trials generally 50% of the subjects are to be administered formulation A in the first period and 50% formulation B to guarantee balanced design. Whether the subjects receive A or B depends on chance.

Statistical evaluation

The proper statistical evaluation of a bioequivalence trial requires some in-depth knowledge of analysis of variance. Explaining all the details of analysis of variance would require a major textbook and is therefore beyond the scope of this review. We must limit the following discussion to some major points.

- Periods (A and B)
- Sequences (A-B and B-A)
- Treatments (formulations A and B)
- Subjects (N)

Fig. 12 Factors of importance in the evaluation of a cross-over trial

A bioequivalence study has four factors to evaluate. The factors are periods, treatments, subjects and sequences. As can be seen, many interactions between these four are possible. There are so many possibilities that interpreting all these interactions become highly complicated and the most often chosen solution is simply to ignore them (and add them to the error mean square). The effects on the power of the trial will be moderate, because it is expected that most of these interactions are non-significant.

When incorporating factors into an analysis of variance-model, some factors are not drawn from a population. When one compares two preparations, it will be evident that the conclusions on equivalence only are valid for these two preparations. One cannot extrapolate the outcome to all preparations. In other words the preparations are not drawn from an infinite number of preparations: they are therefore considered as fixed. A similar line of reasoning is possible for the periods and sequences in a bioequivalence test.

It is easy to see that the findings of a bioequivalence study are to be extrapolated to the entire population, so the subjects must reflect the population: they constitute an unbiased sample. They are drawn from the population and are therefore, by definition, random factors. When incorporating both fixed and random factors into an ANOVA-model, the model is called a mixed model and that has consequences for testing the significance of the four main factors. What the consequences are exactly is again beyond the scope of this review.

A further, rather technical, issue is that according to the FDA/EMEA guidelines on equivalence studies, one has to nest the subjects in the sequence term. Again, beyond the scope of this review.

Analysis of variance is the method of choice for evaluating bio-equivalence studies. A basic assumption underlying the analysis is that all data are normally distributed and there are several reasons to question whether our kinetic parameters are normally distributed. These reasons are:

1. pharmacokinetics: the AUC can be approached as the ratio of the dose times fraction absorbed and the clearance. As one can see this is a ratio and the AUC will therefore tend to be log-normal distributed.

2. statistical: often biological data show a log-normal distribution. AUCs are biological parameters are there is no reason to exempt AUCs from this log-normal principle.

On the basis of the above a ln-transformation of the measured data is now 'obligatory'. Some may remark that tests for a normal distribution of the data exist, but remember that distribution tests do not work well with the relatively small amount of data in a bioequivalence study. The recommendation is therefore always to ln-transform the data. This has however as consequence that when one evaluates the difference between two ln-transformed parameters, one actually looks at their ratio!

Sample sizes

For sample size equations of an equivalence trial one cannot use anymore the classical equations for a cross-over trial as one would in the power approach. Basically the sample size equations of a current type of equivalence study evaluation is necessary for the following reasons:

1. one may never cross the upper or lower limit due to a lack of power, so increasing variability (= wider confidence intervals) must be reflected in larger sample sizes to compensate;

2. larger deviations from unity of the AUC-ratio automatically must lead to larger sample sizes in order to shrink the confidence interval and keep it within the boundaries if the equivalence limit (see above).

It is evident that one has to decide beforehand what the expected variability (or the mean square error form the analysis of variance) is and how large the differences between the formulations will be. One might be optimistic but that frequently means that the sample size will be small and the power of the study low. If they are too low inequivalence is automatically the result, because one cannot rejected the null hypothesis of inequivalence due to a lack of power! Frequently one follows a worst-case scenario and tries to decide what the highest realistic variability will be, coupled to the highest expected difference between the formulations. This can mean that the trial is overpowered, which is good for equivalence testing but can result in significant differences between formulations! These differences are however always smaller than the equivalence limit and therefore of no therapeutic consequence at all.

The equation for the sample size for a bioequivalence study based on ln-transformed data is:

$$N = 2 \times \left[t_{(\alpha, N-2)} - t_{(\beta, N-2)} \right]^2 \left[\frac{CV}{\ln \nabla - \ln \theta} \right]^2$$

Eqn 9 Sample size calculation for a bioequivalence trial based on ln-transformed data.

where the coefficient of variation $CV = \sqrt{e^{MSE^2} - 1}$, the ratio of the means of the two formulations $\theta = \frac{\mu_{test}}{\mu_{ret}}$ and ∇ is the equivalence limit. As one can see the sample size will increase when the variability (expressed as CV) increases and when the ratio of the means deviates from the value 1.
Now this equation has a problem: in order to calculate N (the number of volunteers) one has to known the T-value from the Student T-distribution for N-2 degrees of freedom. However N is unknown, so we can never calculate N. The only solution to this "biting of one's own tail" is trial and error or, better said, an iteration process. Tables have therefore been published (below, one for an equivalence limit of 0.2x μ_{ref}).

CV	*ratio between the means of the measured parameter Θ*						
	0.85	*0.90*	*0.95*	*1*	*1.05*	*1.10*	*1.15*
0.10	36	12	8	6	8	10	20
0.125	54	16	10	8	10	14	30
0.15	78	22	12	10	12	20	42
0.175	104	30	16	14	16	26	56

Equivalence limits

The equivalence limits currently accepted by the regulatory bodies are as follows:

1. for the AUC's 0.8-1.25
2. for C_{max} 0.7-1.43 (although these are currently debated)

These limits are not symmetrical as can be seen. Before one routinely ln-transformed the parameters the limits were respectively 0.8-1.2 and 0.7-1.3. Ln-transformation of these limits result in the following ranges –0.22314 to 0.18232 for the AUC and -0.35667 to 0.26236 for C_{max}. it is clear that the ln-transformed equivalence are not symmetrical any more, so it was decided to adopt the following symmetrical ranges –0.22314 to 0. 22314 for the AUC and -0.35667 to 0. 35667 for C_{max}. Back-transformed these ranges then become 0.8-1.25 and 0.7-1.43 respectively.

Problems with the interpretation of the results

When we carry out the analysis of variance (anova), we are not looking for significant effects for treatment, sequence or periods. The assumptions underlying the cross-over design dictates that these effects are non-significant because they invalidate the trial. So if significant effects are encountered, what do you do? (Please remember that a significant subject effect will always be present. It simply tells you that subjects do differ from each other, which is not a problem. It would be strange if we found that subjects did not show significant differences: it would mean that identical twins or cloned volunteers had been used).

Significant treatment effect

A significant effect for treatment, can simply be ignored. Please remember that a significant treatment effect can be present, when the treatment mean square is small. In other words, remember that the anova procedure we carried out is nothing but the evaluation identical to the power approach, we can understand that significant differences can occur at the moment the variability is low or the number of volunteers sufficiently high. In these terms we have rejected the power approach null hypothesis of equivalence, which is of no interest to us anymore.

The decision of equivalence is based on the Schuirmann test and when the 90% confidence interval is within the equivalence boundaries, there is nothing to worry about. (Basically we just included too many volunteers).

Significant period effect

A significant period effect is caused by the fact that in one of the two periods, the plasmalevels (and AUC) are higher/lower than in the other. The causes may be many: suppose that just before the second period all volunteers received grapefruit juice to drink instead of water. Grapefruit juice inhibits the metabolism of certain drugs. The levels increase therefore in the group receiving A and B. This is not a carry over problem (and because the effect is equal in both groups it is undetectable using carryover tests).

There is still discussion on the meaning of a significant period effect. Some argue that if both treatments are affected in an equal way their relationship (=differences) do not change and hence the comparison between the two is valid. This is basically true but the underlying assumption is that both treatments are indeed equally affected and that has to be proven! Frequently the cause of a period effect is not known and the proof of equal changes becomes difficult.

Significant sequence effect

To gauge the meaning of a sequence effect, one calculates the difference in the AUC's in both sequences (A-B and B-A). Suppose that the A-B sequence difference amounts to -0.5, while the B-A difference is 0.5: the total is zero. Now suppose that the A-B sequence difference amounts to -1.5, while the B-A difference is -0.5: the total is –2, which indicates a sequence effect. The differences between A and B depends on the sequence of the administration. This might be unequal carryover, but it also might be a treatment by period interaction. From the discussion on the cross-over design it is evident that these effects are confounded and cannot be separated. So when a significant sequence difference is present the cause may not be found with certainty. Therefore under special circumstances the significant sequence effect can be ignored. The study must have 1] been a single dose study, 2] been in healthy normal volunteers, 3] not been comparing an endogenous substance, 4] had an adequate washout and 5] used an appropriate design, analysis and equivalence must be present.

An example

To illustrate all the above we list here the results of a true but unrecognisable bioequivalence study.

Manufacturer X has developed a generic equivalent of drug Y and tested the formulation in healthy subjects. Plasmalevels were measured and converted into the AUC_{0-t}. In the table below the resulting AUC, together with the sequence and period order are shown (the raw plasmalevels are not shown, only the AUC values). There were 24 subjects, two treatments (1=ref, 2=test), two sequences (1= test-ref, 2=ref-test).

subject	seq	period	treat	AUC	lnAUC
1	2	2	1	4,409	1,484
2	1	1	1	4,156	1,424
3	2	2	1	4,611	1,528
4	2	2	1	7,163	1,969
5	1	1	1	3,786	1,331
6	1	1	1	6,119	1,811
7	2	2	1	4,097	1,410
8	2	2	1	5,443	1,694
9	1	1	1	5,269	1,662
10	1	1	1	4,472	1,498
11	2	2	1	5,857	1,768
12	2	2	1	6,760	1,911
13	1	1	1	6,097	1,808
14	1	1	1	6,453	1,865
15	1	1	1	5,214	1,651
16	2	2	1	4,812	1,571
17	1	1	1	4,076	1,405
18	2	2	1	6,348	1,848
19	1	1	1	5,587	1,720
20	2	2	1	4,613	1,529
21	1	1	1	6,249	1,832
22	2	2	1	5,906	1,776
23	2	2	1	6,020	1,795
24	1	1	1	7,027	1,950
1	2	1	2	3,867	1,353
2	1	2	2	4,117	1,415
3	2	1	2	5,819	1,761
4	2	1	2	5,774	1,753
5	1	2	2	4,981	1,606
6	1	2	2	5,578	1,719
7	2	1	2	5,460	1,697
8	2	1	2	5,613	1,725
9	1	2	2	4,912	1,592
10	1	2	2	4,644	1,536
11	2	1	2	5,039	1,617
12	2	1	2	5,809	1,759
13	1	2	2	4,746	1,557
14	1	2	2	6,959	1,940
15	1	2	2	5,196	1,648
16	2	1	2	4,982	1,606
17	1	2	2	4,582	1,522
18	2	1	2	6,253	1,833
19	1	2	2	4,320	1,463

subject	seq	period	treat	AUC	lnAUC
20	2	1	2	4,266	1,451
21	1	2	2	4,919	1,593
22	2	1	2	6,784	1,915
23	2	1	2	5,961	1,785
24	1	2	2	6,960	1,940

The data were analysed using SPSS software. The syntax to carry out the proper analysis in SPSS is:

```
GLM
  lnauc  BY period seq treat subject
  /RANDOM = subject
  /CONTRAST (treat)=Simple
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /CRITERIA = ALPHA(.1)
  /DESIGN = period seq treat subject within seq.
```

This syntax instructs to evaluate the ln-transformed AUC values, while the factors included in this model are periods, sequences, treatments and subjects. The factor subjects is random and the others are fixed. A difference between the treatments is calculated (=contrast statement) and SPSS automatically includes the 90% confidence interval of that difference (criteria statement). The design statement indicates that the subjects are nested within the sequences.

This syntax commands results in the analysis of variance table below:

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
PERIOD	Hypothesis	3,324E-03	1	$3{,}324E^{-03}$	,264	,613
	Error	,277	22	$1{,}260E^{-02}$ [b]		
SEQ	Hypothesis	2,299E-02	1	$2{,}299E^{-02}$	,449	,510
	Error	1,127	22	$5{,}123E^{-02}$ [a]		
TREAT	Hypothesis	4,328E-03	1	$4{,}328E^{-03}$	,344	,564
	Error	,277	22	$1{,}260E^{-02}$ [b]		
SUBJECT(SEQ)	Hypothesis	1,127	22	$5{,}123E^{-02}$	4,066	,001
	Error	,277	22	$1{,}260E^{-02}$ [b]		

[a] tested against mean square of subjects nested within sequences
[b] tested against mean square error

From this table it can be seen that the period, sequence and treatment effects are non-significant and (as expected) the subjects do differ. Also, it is indicated that the F-test for the factors uses different denominators: for the period, treatment and subjects within sequence, the mean square error is used to calculate the F-value, but for the sequence term it is the 'subjects within sequence' mean square. This is related to the fact that the subjects are considered to be a random factor.
SPSS was also instructed to calculate the 90% confidence interval of the difference between the two ln-transformed AUC means.

		LNAUC
Contrast Estimate		1,899E^{-02}
Std. Error		,032
Sig.		,564
90% Confidence Interval for Difference	Lower Bound	-3,665E^{-02}
	Upper Bound	7,463E^{-02}

The important conclusion is that the 90% confidence interval ranges from -3.665×10^{-2} to 7.463×10^{-2}, while the mean difference is 1.899×10^{-2}. Back-transformed to regular units this means that the mean AUC-ratio is $e^{1.899 \times 10^{-2}} = 1.01917$, while the 90% confidence interval ranges from $e^{-3.665 \times 10^{-2}} = 0.96401$ to $e^{7.463 \times 10^{-2}} = 1.07749$. This makes an equivalence claim possible.

Suggested reading

Books

- *on bioequivalence testing*
  Chow and Liu: Design and analysis of bio-availability and bio-equivalence studies.
  Marcel Dekker

- *on the statistical background, analysis of variance and the cross-over trial design*
  Kirk RE: Experimental design, procedures for the behavioural sciences.
  Brooks/Cole Publishing Company 3rd ed.

Senn S: Cross-over trials in clinical research.
John Wiley

Journal articles

- *the classical publication*
  Schuirmann DJ. A comparison of the two one sided tests procedure and power approach for assessing the equivalence of average bioavailability.
  J Pharmacokin Biopharm 15: 657-680 1987

- *on the power of the Schuirmann approach*
  Philips KF. Power of the tow one-sided tests procedure.
  J Pharmacokin Biopharm 18: 137-144 1990

- *on sample size estimation*
  Dilleti E. Hauske D. and Steinijans VW. Sample size determinations for bioequivalence assessment by means of confidence intervals.
  Clin Pharm Ther Tox 29: 1-8 1991

  Hauske D. Steinijans VW. Dilleti E. Burke M. Sample size determinations for bioequivalence assessment using a multiplicative model.
  J Pharmacokin Biopharm 20: 557-561 1992

- *on the general background of pharmacokinetics*
  Riegelman S. Collier P. The application of statistical moment theory to the evaluation of in-vivo dissolution time and absorption time.
  J Pharmacokin Biopharm 8: 509-534 1980

WWW

Many (updated) guidelines on bioequivalence testing can be found on the Web sites of the FDA (www.fda.gov) and the European EMEA (www.eudra.org)