The Swedish Association of Medical Statistics (FMS) and The Danish Society for Biopharmaceutical Statistics (DSBS) organize jointly a one-day meeting in Malmö on the 6th of May 2004 9.15 am – 5 pm on

Statistical Issues in Drug Development

Key note speaker: Professor Stephen Senn

More information about the meeting and how to register can be found at http://www.statistikersamfundet.se/fms/

Below you will find abstracts for the talks.

Added Values

Stephen Senn
Department of Statistics University of Glasgow

"As ye randomise so shall ye analyse," is one way of describing Fisher's defence of randomisation. Yet, when it comes to clinical trials we nearly always randomise but we rarely analyse the way we randomise and Fisher himself was no exception. Two controversies involving Fisher in the 1930s are discussed: one with Neyman concerning additivity and the other with Student concerning randomisation. Their relevance today is considered, as is whether randomisation inference in clinical trials is dead and whether modelling rules the day, whether minimisation is an acceptable procedure and to what extent trialists confuse experiments with surveys. It will be maintained that a number of different possible purposes of clinical trials have been confused because in the case of the general linear model, under strong additivity, they can all be satisfied by a single analysis. More generally, however, this is not the case. The field of binary equivalence is an area that brings these conflicts to a head and will be briefly considered. It may be that more explicit declaration of the purpose of an analysis would be helpful in avoiding confusion.

How much shyster do you want with your quack?

Stephen Senn
Department of Statistics, University of Glasgow

The pharmaceutical industry is currently facing a barrage of litigation concerning side-effects of pharmaceuticals (Senn, 2003). In judging individual cases, a key role of statistics appears to be whether or not an adverse event that occurred to a given patient was caused by a pharmaceutical that was taken. Questions of this sort have been referred to as a "causes of effects" questions: the "effect" has been observed, it is desired to establish the "cause"(Holland, 1986). The chronological logic is similar to that which applies to a case-control study. However, traditional analysis of clinical trials concerns itself with "effects of causes" questions: a patient is given an experimental treatment, a putative "cause" and its "effects" are studied. In litigation concerning pharmaceuticals, a key role has been given to odds ratios. In particular it has been argued that where it has been established that the general odds ratios for an adverse event for a given treatment compared to placebo is in excess of two, it is then more likely than not, that where such an event occurs, and the treatment has been taken, the treatment is the cause. This would seem to open the door for the methodology of equivalence testing to be applied to demonstrate the safety of pharmaceuticals: one needs to show that the odds ratio for any adverse event is less than two.

It will be argued here, that a programme of such proof, whether in a Bayesian or frequentist framework, is both impossible(Senn, 1993) and irrelevant. Taking various concrete illustrations, it is shown that contrary to initial appearances, it is the effect of causes question that remains relevant. The implications of this for the analysis of safety profiles of pharmaceuticals is considered.

References:

Holland, P. W. (1986). *Statistics and Causal Inference*.Journal of the American Statistical Association 81(396): 945-960.
Senn, S. J. (1993). *Inherent difficulties with active control equivalence studies*.Statistics in Medicine 12(24): 2367-75.
Senn, S.J (2003) *Dicing with Death*, Cambridge University Press, Cambridge
Senn, S.J. (2004) *How much shyster do you want with your quack.?*

Examples of multivariate dose-potency estimation

Anders Källén
AstraZeneca R&D, Lund, Sweden

As a follow-up to a previous publication we illustrate how multivariate analysis of dose-response trials can be done, providing single number, relative dose potency, results for the comparison between two treatments, also when a number of variables are measured.

Quantifying the within-subject variation of insulin action profiles

Lars A. Endahl and Birgitte B. Rønn
Biostatistics, Novo Nordisk A/S, Denmark

Patients with Type 1 diabetes have lost the ability to produce insulin and consequently rely on daily injections of insulin in order to stabilise blood glucose concentration. As both too high blood glucose values (hyperglycaemia) and too low values (hypoglycaemia) are potentially life-threatening, the day-to-day variability of the pharmacodynamic profile following insulin injection is of utmost importance. Trial NN304-1450 was a four-period parallel-group trial encompassing 54 subjects with Type 1 diabetes with the primary objective to assess the within-subject variability of the pharmacodynamic profiles of the novel insulin analogue, insulin detemir, in comparison with that of the most commonly used basal insulin preparations, NPH insulin and insulin glargine. Pharmacodynamic action was assessed as the amount glucose infused in order to maintain a constant blood glucose concentration following insulin injection. The average and maximum effect were derived from the pharmacodynamic profiles and analysed after log-transformation using an ANOVA model with insulin preparation and treatment period as fixed effects; subject as a random effect with a variance dependent on insulin preparation and an error term with a variance dependent on insulin preparation. The error term variances were square-root transformed and reported as within-subject CVs. The usual prediction intervals, based on the population variation does not illustrate the important variability for the patients as insulin therapy is individually adjusted. Thus in order to illustrate the within-subject variability we employed 95% conditional prediction intervals, which by definition display 95% of the predicted values for a randomly selected subject. The prediction intervals were then used for clinical interpretation of the differences in the expected frequencies of hyper- and hypoglycaemia for the three insulin preparations. *Acknowledgement:* The design and analysis of trial NN304-1450 were carried out in collaboration with Eberhard Draeger, Tim Heise and his staff from Profil Institut, Neuss, Germany.

References:
Heise T, Nosek L, Rønn BB, Endahl L, Heinemann L, Kapitza C, Draeger E (2004): *Lower Within-Subject Variability of Insulin Detemir in Comparison to NPH Insulin and Insulin Glargine in People with Type 1 Diabetes*. Diabetes, In press.

A comparison of multiplicity-adjustment methods in a dose-finding study with a binary endpoint.

Nedjad Losic,
SpadilleMediMentum A/S, Denmark.

In order to maintain the overall type I error in a dose-finding study, multiplicity adjustment is typically necessary. The performance of five different adjustment procedures on some different dose-response relationships where the response variable is binary, are compared with respect to significance level and power. The discreteness and correlation in the data are taken into consideration. The results, based on simulations,

show that it is possible to achieve nearly the same power as with no adjustment. Hints regarding implementation in SAS are given.

Interim analyses and other early looks at trial data: Contribution or curse?

Kim Mark Knudsen
Genmab & Kristian Windfeld, Ferring Pharmaceuticals, Denmark

In drug development, we are often requested to present results from clinical trials even before the trial is completed. From these early looks at data, unduly enthusiastic or pessimistic recommendations may follow based on interim results in order to reduce development time and costs, often in non-blinded trials. More than just the additional labor connected with conducting the analysis, we find ourselves in a dilemma as to the soundness of this practice. We will try to argue both ways: for a controlled use of interim monitoring of accumulating data using various statistical approaches and data monitoring committees; and against numerous resource-demanding analyses that can put the trials' credibility and integrity at risk and can in the worst case jeopardize the clinical project. We are not even sure that the two speakers can reach a consensus.

Statistical significance and clinical relevance

Olivier Guilbaud
AstraZeneca R&D, Södertälje, Sweden

Given an observed difference between two treatments means that is statistically significantly larger than zero in a clinical study, it may be intuitively reasonable for an "interpreter" to look at the magnitude of the point estimate, i.e. of the observed mean difference, to make a judgment about the clinical relevance of this finding. The following aspects related to such an interpretational behavior are considered in this presentation: (a) which types of errors can be made; (b) how large are error probabilities; and (c) how can a study be designed so that the power is high of getting a "desirable outcome". Here a "desirable outcome" consists of an observed mean difference in the study that is both statistically significantly larger than zero and larger than a positive bound anticipated to be considered as being clinically relevant and/or important by "interpreters". The consequence of determining sample sizes to merely "detect" a certain difference with high power is discussed in this context.

Non-inferiority testing in clinical trials: Concerns regarding credibility

Lars Frison
AstraZeneca R&D, Mölndal, Sweden

Clinical studies aiming at establishing non-inferiority are getting increasingly more common. There is no consensus on how scale and size for non-inferiority criteria (delta)

should be determined. Regulatory-based studies are usually “forced” to be reasonably credible, while publication-based studies are all over the place. There are many bad examples, clinical trials may lose credibility.

In general terms it is defined in regulatory guidelines, e.g. ICH E10, how a delta should be defined. Firstly, the margin chosen can be no greater than the smallest effect size that the comparator would be reliably expected to have compared with placebo in the setting of the planned trial. Secondly, the margin has to be smaller than differences between treatments regarded as being clinically significant. This last criteria is subjective, requires medical judgment and can only make the delta smaller relative to the first criteria.

For a non-inferiority study to be credible the delta has to be chosen such that a substantial proportion (e.g. 50%) of the comparators efficacy advantage is retained. However, this is not enough, it also has to be established, as well as possible, that the current study possesses assay sensitivity, and that constancy applies for the benefit of the comparator relative to placebo. Those critical aspects of a non-inferiority design will be discussed.

The choice of scale for testing non-inferiority is important. Both absolute and relative comparisons are used, and tests are performed with or without transforming data. The relative merits of the various approaches will be addressed and circumstances when concerns may be expected will be elucidated.