Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 544 (2005) 268–279
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

The variance of screening and supersaturated design results as a measure for method robustness

B. Dejaegher, J. Smeyers-Verbeke, Y. Vander Heyden*

Department of Pharmaceutical and Biomedical Analysis, Pharmaceutical Institute, Vrije Universiteit Brussel VUB, Laarbeeklaan 103, 1090 Brussels, Belgium

Received 11 October 2004; received in revised form 20 December 2004; accepted 20 December 2004
Available online 19 January 2005

Abstract

Screening designs are factorial designs to evaluate the importance of factors in a number of experiments that is at least one higher than the number of factors examined. Supersaturated designs are factorial designs with more factors than experiments. These designs do not allow a correct estimation of the factor effects due to a confounding of main effects. Therefore, it is evaluated whether, in robustness testing, the variance of a response can be used as a measure for the robustness of a method. A number of potential reference criteria (reference variances estimating reproducibility and limit values) also are evaluated for their applicability to decide whether the examined factors cause non-robustness. Finally, it was also examined which conclusions one statistically can draw from comparing the variances from the design experiments with the reference criteria. Two approaches are considered for the reference variances: a classical *F*-test and interval hypothesis testing. The use of some limit values was also discussed. It was found that the variance of a response could be used as a measure for robustness, but statistically this variance could not be interpreted in an acceptable way. Either a large probability to accept a non-robust or to reject a robust method occurs due to the small number of degrees of freedom to examine a given number of factors, especially when applying supersaturated designs.

Keywords: Screening designs; Supersaturated designs; Variance; Robustness; *F*-test; Interval hypothesis testing

1. Introduction

Screening designs, such as Plackett–Burman (PB) or fractional factorial designs, are usually applied in robustness tests [1–3]. Factors, responsible for non-robustness of a method, show the largest change in response for a small change in their levels. A method is considered robust when its quantitative aspect (e.g. content determination) is not influenced by the factors examined [2].

A full factorial design examines all possible combinations between the different factors and their levels. It allows calculating all main and interaction effects. The number of experiments required increases exponentially with the number of factors [1]. Usually for a given method a large number of potentially relevant factors can be selected, but only a few are believed to be important. This is called the effect sparsity principle [2,4,5]. If the aim is to detect those few active factors, usually a fractional factorial design is performed [1–3,6]. In these designs a confounding of effects occurs, i.e. main effects cannot be estimated separately from some interactions. A saturated fractional factorial design is the smallest fraction of a full factorial in which the main effects still can be estimated, unconfounded from each other. PB designs are factorial designs that examine up to $N - 1$ factors in N (multiple of four) experiments [1,7]. Main effects are again confounded with interaction effects. When applying these screening designs one assumes that two-factor and higher-order interaction effects are negligible, compared to the significant main effects. A disadvantage from a practical point of view is that for a large number of factors, still rather many experiments are required.

* Corresponding author. Tel.: +32 2 477 47 34; fax: +32 2 477 47 35.
E-mail address: yvanvdh@vub.ac.be (Y.V. Heyden).

0003-2670/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2004.12.056

Therefore, a tendency exist to develop designs that, compared to screening designs, allow examining either more factors in the same number of experiments, or the same number of factors in less experiments. Supersaturated designs (SS), for instance, examine more than $N-1$ factors in N experiments [8,10–12]. However, these SS designs have as disadvantage that no individual main effect can be estimated reliably, since they are confounded. It is thus impossible to identify the factors, responsible for the non-robustness of the method, straightforwardly. These designs anyway potentially can be used in robustness tests, since estimates of the individual factor effects are not necessarily required. The ICH (International Conference on Harmonisation of Technical Requirements for the Registration of Pharmaceuticals for Human Use) guidelines, for instance, defines robustness as: "*The robustness of an analytical procedure is a measure of its capacity to remain unaffected by small, but deliberate variations in method parameters and provides an indication of its reliability during normal usage*" [9]. Therefore, the variance of the response measured from the experiments of a design potentially could be used as a measure for the robustness of the method and this variance can be considered an estimate of the reproducibility of the method [8].

Prior to using SS designs in robustness tests, some questions need to be answered positively: (A) "Can the variance of a response, estimated from a SS design, be used as a measure for the robustness of the method, i.e. does the variance represent the variability introduced by the examined factors?", (B) "Is it possible to define a useful criterion to decide whether the factors examined in the SS design cause non-robustness of the method? Such criterion then also can be used for screening designs, such as PB and fractional factorial designs.", (C) "If such a criterion indeed is available, can statistically relevant conclusions be drawn from its comparison with the variance of the response?" The statistical implications of using the variance from design results to draw conclusions about the robustness of a method thus deserves some thorough examination and was tackled.

The first question has been examined earlier [8]. The outcome was that the variance estimated from a SS design ($N=6$, $f=10$) is similar to the one from a PB design ($N=12$, $f=11$). Since the latter is appropriate to estimate the factor effects, its variance can be considered a measure for the variability caused by the variation in the examined factors. Therefore, the variance estimated from a SS design also is. The two remaining questions are examined in this paper. They are considered both for screening and supersaturated designs. Some criteria (reference variances or limit values) to draw conclusions about the robustness of a method are proposed. To evaluate the proposed reference criteria on their usefulness, both robust and non-robust methods are considered.

A robustness test on a robust method that assays a main and two related compounds, is taken from [3]. Introducing significant effects created simulated robustness test results describing non-robust methods. In both situations the robustness test was performed using a PB design. Supersaturated designs examining the same factors were constructed as in Ref. [8], i.e. using dummy factor columns from the PB designs as branching columns (see Section 2). The similarity of the variances from the PB design and the corresponding SS designs was verified. The proposed reference criteria are calculated. The reference variances are compared with the variance of a robustness test design, following two approaches, F-testing and interval hypothesis testing. For each comparison the α- and β-errors are determined in order to try to answer question (C). The usefulness of the limit values is also evaluated.

2. Theory

2.1. Supersaturated (SS) designs

SS designs can be constructed in several ways [10–12]. Some methods apply a specific criterion in order to approach orthogonality as much as possible [10,11]. Other SS designs are constructed via half-fractions of Hadamard matrices or Plackett–Burman designs [12]. A PB design with N experiments and $N-1$ factors can be split in two SS designs with each $N-2$ factors and $N/2$ experiments. In this paper, we use the latter approach. In the PB design, one column is defined as branching column. Selecting all experiments for which the branching column is either at (−) or (+) level and then deleting the branching column yields both SS designs. An example is given in Ref. [8]. Any column can be used as branching column to construct a SS design. However, the method is not applicable to all PB designs. Exceptions are given in [12].

2.2. Some potential reference variances and limit values

It has been shown that the variance of a response due to the variations introduced in the examined factors may acceptably be estimated from a SS design [8]. As mentioned in Section 1, a reference criterion is needed to draw conclusions about the robustness of the method based on the variance estimated from the design. Such criterion then also could be used to evaluate the (non)-robustness of a method from screening design (PB or fractional factorial designs) results prior to the examination of effects.

The variance, s^2_{design}, calculated from a screening or SS design in robustness testing can be considered an estimate of the reproducibility variance, s^2_{R}, of the method [8]. Therefore, a reference variance that also estimates reproducibility could be applied as possible criterion. However, an appropriate estimate of reproducibility from an interlaboratory study is not available since a robustness test is performed prior to such a study [13]. Variance estimates that already may be available when a robustness test is performed are the repeatability or an intermediate-precision estimate [14]. Horwitz et al. [15] and Boyer et al. [16] made a prediction for the reproducibility via the Horwitz function, which relates reproducibility