Journal of Pharmaceutical and Biomedical Analysis
14 (1996) 1125-1132

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

Robustness testing in liquid chromatography and capillary electrophoresis[1]

H. Fabre

Laboratoire de Chimie Analytique, Faculté de Pharmacie, 34060 Montpellier Cédex, France

Received for review 10 October 1995; revised manuscript received 22 January 1996

Abstract

The definition and objectives of robustness testing are given and the essential features of the methodology which can be applied using a multivariate approach in liquid chromatography and capillary electrophoresis are described. Guidelines are given for the different steps which are involved in using screening and response surface designs. It is shown that screening designs may be sufficient to set the method limits but that response surface designs are of major interest in method transfer because they give a comprehensive picture of the behaviour and limitations of the method.

Keywords: Capillary electrophoresis; Liquid chromatography; Response surface designs; Robustness testing; Screening designs

1. Introduction

Following the last International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), a tripartite guideline on the validation of analytical procedures was published for registration applications within the European Community, Japan and the United States [1]. The most important criteria listed for validating different types of analytical procedures do not differ fundamentally from those given in the USP XXIII [2] and European Explicative Note [3] but a new terminology is proposed concerning precision, reproducibility and robustness. It should be noted that the latter is not listed as a typical validation criterion, although it is defined in the attached glossary, because it should be "... considered at an appropriate stage in the development of the analytical procedure" [1].

This paper is intended to provide the analyst, via selected experimental designs with practical guidelines in the methodology for robustness testing in liquid chromatography (LC) and capillary electrophoresis (CE), which are both characterised by a large number of factors likely to affect the separation and quantitative results. These guidelines are based on laboratory experience of the author on robustness testing in high performance

[1] Presented at the Fifth International Symposium on Drug Analysis, September 1995, Leuven, Belgium.

PII S0731-7085(96)01770-0

liquid chromatography (HPLC; [4,5]), thin-layer chromatography (TLC; [6]) and a series of papers relating to HPLC [7–15] and CE [16,17]. For comprehensive information on experimental designs, the reader should refer to the specialised literature (see e.g. Refs. [18–25]).

The definition and objectives of robustness testing are given and the essential features of the methodology which can be applied in LC and CE are described.

2. What is robustness?

As defined by the ICH, the robustness of an analytical procedure refers to its capability to remain unaffected by small but deliberate variations in the method parameters [1]. It is distinct from the precision, which covers the terms repeatability, intermediate precision and reproducibility. Repeatability (or intra-assay precision) expresses the precision under the same operating conditions over a short interval of time, intermediate precision refers to within-laboratory variations (different days, analysts, equipment, ...), while reproducibility refers to between-laboratory variations. These definitions are very close to those presently discussed by the International Organization for Standardization [26]. It should be noted that the terminology ruggedness in the USP XXIII [2] corresponds to intermediate precision, reproducibility and robustness as a whole. At least, the ruggedness testing in chemometrics and statistics [19] refers to the ICH definition of robustness testing.

3. Why robustness testing?

Robustness testing identifies the factors in the method which have a significant effect on its results and anticipates the problems which may arise during its application on different instruments, using different reagents, batches of thin layer chromatographic plates, chromatographic columns or capillaries, and in different environments. It gives an indication of the intermediate precision and reproducibility which can be expected from the method because it evaluates its capacity to withstand changes being encountered in different laboratories. It allows limits to be set for all method parameters and, where they are very narrow, to underline in the protocol the limits permitted. Therefore, it provides useful information for method transfer.

4. Methodology

A multivariate approach in which simultaneous changes are made in the method factors using a matrix of experiments is recommended. However, since robustness is generally investigated when the method has already been optimised and is aimed at specifying factor limitations, the experimental approach used is different from that of optimisation. The levels of variation selected for the critical factors are small, which allows screening purpose fractional factorial designs to be used, assuming high-order, third-order and even second-order interactions are negligible. The most important factors having been identified, they can be tested with a more informative design such as a Box-Behnken design or a central composite design which will give a complete description of the system in the method region including factor interactions and squared terms. The response surface allows the method factor limits to be set.

For both categories of designs (screening designs and response surface designs), the robustness test involves five steps:
—selection of the critical factors,
—selection of the factor levels and level number,
—selection of an experimental design,
—realisation of experiments,
—statistical analysis of the responses and interpretation.

These steps will be considered successively for the two classes of designs mentioned above.

4.1. Screening designs

The goal here is essentially to identify critical factors in the analytical procedure.

4.1.1. Selection of the critical factors

Critical factors are quantitative or qualitative factors which are believed to affect the results. Some 50 have been listed as being liable to influence HPLC methods with UV detection [10]. Further factors can be listed if the method includes pre- or post-column derivatisation [5]. These factors are related to the sample preparation step (sample weight, internal standard concentration, sonication time, volume of extraction solvent, age of the solutions, etc.), separation and detection. Typical factors are for example the flow rate and mobile phase composition (buffer pH and concentration, smallest component in solvent mix, additive concentration, etc.) in HPLC, the pH, nature of electrolyte or electrolyte concentration, the applied voltage and separation temperature in CE, the tank saturation and development temperature in TLC. Injected or applied volumes (HPLC, TLC), injection times (CE), batch or age of column (HPLC), plate (TLC) or capillary (CE), and detection wavelength are also liable to affect the responses. However, since the changes in the level values are small only a limited number of factors are expected to affect the results. The number of factors selected will depend also on the intended application: internal use, use on different sites, collaborative studies, official methods, in the order listed, require an increasing number of factors to be tested. The number of factors tested is often limited to eight for practical reasons and mainly because of time considerations. HPLC and CE experiments are carried out sequentially (not at the same time) and it is important to perform the experiments over a limited period of time to have reliable results. The time needed for a robustness experiment can be a serious limitation: in HPLC it comprises not only the run time but also the time needed for system equilibration after a change in temperature or solvent composition. Serious limitations may also arise in HPLC for impurity testing around the limit of quantitation which are often due to the problem of column overloading in the presence of a high concentration of the main compound and the need for replicated injections at low concentration levels of degradation compounds. Modern PC-controlled HPLC hardware is now available to facilitate automated programming. Robustness testing of TLC has the advantage that several plates can be developed at the same time. However, ruggedness testing for an impurity determination at the limit of quantitation is difficult [6] due to the limited number of loadings which can be applied on the same plate and the lack of automation. Factors which may be assumed to be less critical (such as on-plate stability or plate-to-plate variations within the same batch) may be tested at early stages of validation and eliminated in the experimental design to reduce the number of experiments. Robustness testing in CE [16,17] is easier to perform in comparison with HPLC and TLC as changes in the electrolyte composition or temperature separation do not require a long equilibration time and complete automation is possible with commercial instruments.

4.1.2. Factor levels

The variations in the factor levels should reflect those which could be encountered in different laboratories due to the use of different instruments, stationary phases, environmental conditions, analysts etc. The number of levels tested for each factor is preferably three (low, high and nominal levels). It should be borne in mind that two-level factorial design implies a linear relationship between the factor and the response, which is not always verified; because the method has already been optimised, the nominal level for one or several factors may be close to the optimum, yielding a non-linear response between the two extreme levels tested apart from this value. The main effect of wavelength will not be disclosed by comparing the two extreme levels if the two wavelengths are situated on each side of the maximum absorbance wave length. The comparison of the responses at low and high levels [13] without running experiments at the nominal level (center point) should not be the rule.

4.1.3. Choice of a screening design

Full factorial designs are not employed for screening purposes due to the large number of experiments involved. In full factorial designs, the number of experiments ($N = l^n$) corresponds to all possible combinations of selected factors (n) and levels (l), which means that for eight factors tested at two levels, 256 experiments are required. Fractional factorial designs which use a fraction of the full factorial design and consider that high-order, second-order or third-order interaction effects are negligible are highly economical for screening. The selection of an appropriate fractionation is based on several considerations, namely the number of critical factors (often six to eight) and levels (two or three), and the number of experiments which can be carried out in a reasonable period of time. This is highly dependent on the degree of automation. It should be noted that if the experiments cannot be carried out in one run, it is possible to block the design but some effects are confounded and information is lost.

Highly-fractionated designs such as saturated fractional designs which represent the highest degree of fractionation possible are useful for screening a high number of factors. Plackett and Burman designs [24] are a category of saturated designs which have been proposed for the robustness testing of official methods [27] and have often been used in LC [4,5,7–12,14]. In these designs the number of experiments required is equal to the number of factors $n+1$ and is a multiple of four for a factorial design with two levels. Plackett and Burman designs with seven and 11 factors are useful for LC and CE. The effect of the factors has been investigated generally at three levels using a reflected saturated design which requires 15 experiments for seven factors, the experiment at the nominal level being common for the high and low levels; the effect at both levels is evaluated by reference to the nominal level. These designs estimate independently the main effect of each factor but have the drawback of presenting a severe confounding pattern [9,24]. For example, for seven factors tested at two levels, each main effect is confounded (aliased) with three two-factor interactions. Therefore, a main effect will be real only if the interactions are insignificant. As factors which have a large effect are more likely to produce interactions with other factors, the confounding pattern has to be carefully examined before planning the experiments whilst attributing a place to a factor. For a number of factors lower than seven, or for eight to ten factors, Plackett and Burman designs do not exist and one or several dummy variables (imaginary factors) are used as factors to complete the existing designs. For example, eight factors at two levels can be tested with a 11 factor factorial design with three dummy variables. These dummy variables do not represent a real change and can be used in statistics to evaluate the repeatability of the procedure [8,9,14]. However, if one considers a Plackett and Burman design which is a power of two, it is recommended to choose an equivalent saturated design from the two-level factorial menu as this gives the best possible alias structure for the combination of factors and experiments selected.

If two- or higher-factor interactions are suspected, fractional designs other than saturated designs should be used in order to separate main effects from interactions, which require a higher number of experiments. The degree of fractionation selected depends on the number of experiments allowed. For the screening of eight factors at two levels in CE [16,17], fractional designs 1/8 and 1/16 of full factorial designs with the addition of four center points have been used which require 36 and 20 experiments respectively. Inclusion and replication of a center point in random order throughout the set of experiments is used to evaluate the repeatability of the procedure and test the curvature. If the design is blocked, each block should comprise one or several center points. Such designs are less ambiguous than the saturated designs for the evaluation of main effects because of their higher resolution.

4.1.4. *Realisation of experiments*

Solutions: for the evaluation of the critical chromatographic or electrophoretic parameters, a mixed standard solution is prepared. For an assay, standard and test solutions are prepared and injected in the sequence indicated in the procedure.

Injections: replicate injections should be preferred, except if the time is restricted, to estimate the repeatability; experimental designs are not often replicated.

Order of experiments: it has been stated that randomisation is not necessary when one is primarily concerned with using screening designs with one observation per run [24]; however, it is the opinion of the author that the different experiments should preferably be carried out in a random order (selected from a table of random order or generated by software) to take into account uncontrolled factors likely to introduce a bias into the responses. Randomisation is essential if a center point is used.

Responses: quantitative responses (peak areas, peak heights) but also qualitative chromatographic or electrophoretic parameters (plate count, symmetry factor, resolution, retention or migration times) are typically measured during the tests.

4.1.5. Statistical analysis of the responses

The result of experiments are submitted to a statistical analysis. Calculations can easily be done manually or with appropriate software if an ANOVA test is carried out.

(i) The effect of each factor on the response can be manually calculated by the difference D between the mean values of the responses obtained at both levels. In the Youden and Steiner approach [27], the significance of this difference is evaluated by a Student test using the standard deviation (SD) calculated from the results of experiments. A better and more severe approach uses the SD calculated from replicate determinations at the nominal level performed during method validation [5,6,19] or preferably throughout the test. In the method of Box et al., [18] exploited by Mulholland and co-worker [8,9] on a reflected Plackett and Burman design, the main effect (ME) for each factor is calculated by the difference between the mean response at the nominal and extreme values normalised to the response obtained in the experiment at the nominal value and expressed as a percentage. The value of the ME for each factor can be compared to that obtained for the dummy variable which reflects the variability of the procedure [8,9,12,14].

(ii) It is also possible using classical statistical software to estimate each effect by a multiple regression fitting of a mathematical function: $Y = b_0 + b_1A + b_2B + b_3C + \cdots$, for a model without interaction, $Y = b_0 + b_1A + b_2B + b_3C + \cdots, + b_{12}AB + b_{13}AC + \cdots$, for a model with interactions, where Y is the experimental response obtained with the factors A, B, $C, \ldots$ at level $\pm$, the regression coefficients b_1, b_2 are the main effects of factors A, B, and the regression coefficients b_{12}, b_{13} are two-factor interaction effects.

For example, the different steps in the analysis of a factorial design at two levels (high and low) with a replicated center point are as follows: (i) selection of a mathematical model; (ii) fitting of the results with the selected model (using the replicated center point to calculate the pure error); (iii) evaluation of the curvature (using the data of the center point) to check the goodness-of-fit of the planar two-level factorial model (curvature of the surface may indicate that the design is in the region of an optimum); (iii) calculation of the significance of each factor using a Student test (t value = regression coefficient/pure error).

The statistical results can be supplemented by graphic plots of the main effects and dummy variable effect [9], normal probability plots which visualize the critical factors [23], interaction plots and Pareto plots which give rapid visual information on the size of the effects [16,17].

4.1.6. Conclusion from screening experiments

The identification of critical factors in screening experiments will point out the necessity of their control in order to avoid a drift in the application life of the method. However, factors can influence the chromatographic or electrophoretic parameters, or the assay results, but the responses can still be within the acceptance criteria. If the responses obtained for each experiment comply with the method requirements, a screening design can be sufficient to set the method limits, allowing adequate system suitability tests and assay, at the extreme levels used in the robustness test. However, if some factors are shown to have large effects on the responses and if the results are outside the limits specified, further experiments may be carried out using response surface designs [16,17] to explore the response as a function of one or several factors around the method values.

4.2. Response surface designs

The goal here is to determine the specification limits and predict the variation of the response (resolution, asymmetry, retention or migration time, etc.) inside or slightly outside the area investigated in screening experiments.

4.2.1. Selection of the critical factors

The factors to be tested are those which are already known or have been found to produce

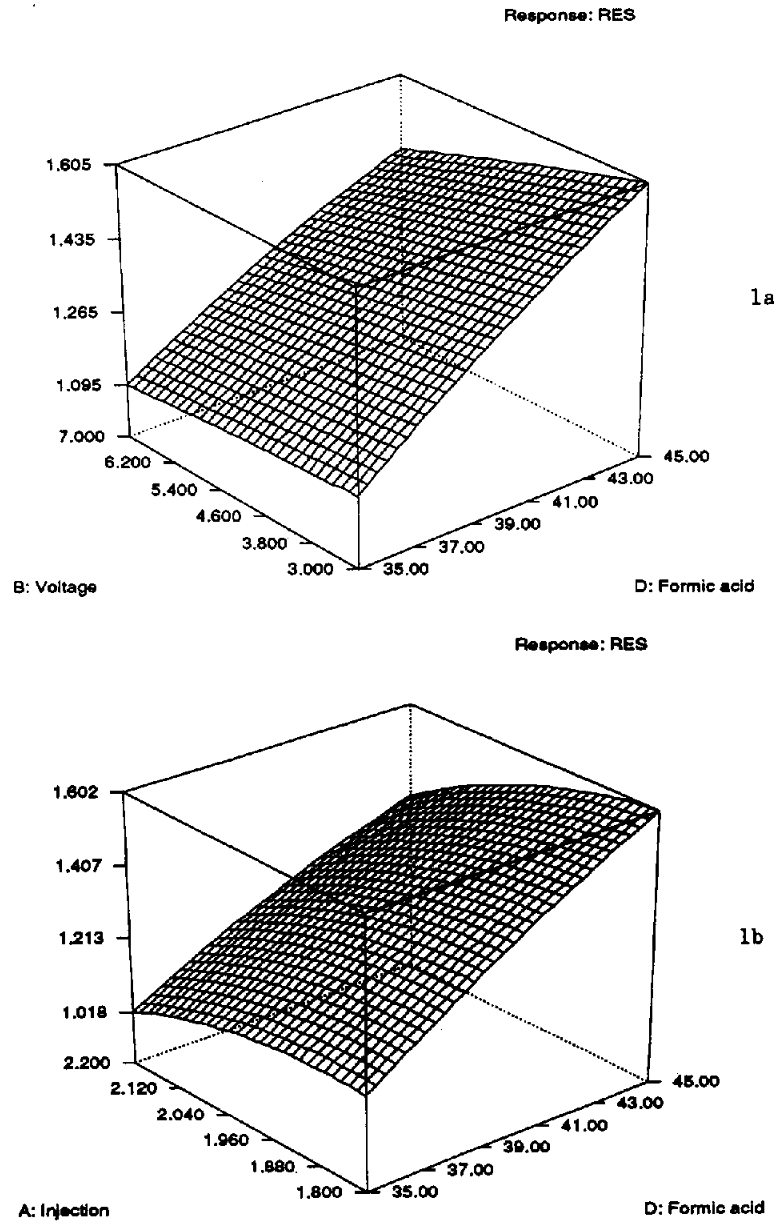

Fig. 1. Response surfaces obtained for resolution between sodium and potassium in a CE method: (a) resolution as a function of voltage and formic acid concentration (from Ref. [17]); (b) resolution as a function of injection time and formic acid concentration (from Ref. [28]).

large effects in screening experiments and generally amount to between two and four.

4.2.2. *Factor levels*

A minimum of three levels for each factor is needed to describe a response surface, but five levels are preferred for a more accurate surface response.

4.2.3. *Choice of a design for response surface*

Response surfaces can be obtained only with designs in which the two- or three-factor interactions are taken into account. Composite designs are very efficient for a limited number of factors ($<5-6$) and are therefore well suited for specifying factor tolerances [19,20,23]. They have been used in conjunction with fractional factorial designs as screening designs to set suitability criteria in CE [16,17]. Central composite designs consist of the juxtaposition of a star design with $2n$ axial points and a factorial design with 2^n factor combinations (or a fractional factorial design), augmented by one center point minimum; the centers of the two designs coincide. The replication of the centroid is particularly important as its response may dramatically affect the shape and orientation of the response surface. The number of replicates depends on the size of the design and the requirements of the property searched (orthogonality, rotatability, etc.). Experimental central composite matrices which have been used in CE can be found in Refs. [16,17]. Other designs are proposed by the software for evaluating response surfaces.

4.2.4. *Realisation of experiments*

The guidelines indicated for screening designs are applicable. Randomisation is essential for all the experiments.

4.2.5. *Statistical analysis of the responses*

The matrices are computed. An ANOVA test is used to test the fit of the data to a selected model. A quadratic model which also evaluates two-factor interaction effects is generally adequate in most cases in CE and HPLC [28]. The response surface plot allows visualization of the variations of the response as a function of the level of the factors. Examples of response surface plots without and with interaction for the resolution between potassium and sodium as a function of voltage (or injection time) and formic acid concentration in a CE method are shown in Fig. 1 [17,28]. The resolution increased with acid formic concentration but voltage had no effect (Fig. 1a). There was no curvature in the response surface because there was no significant interaction between the two factors on the resolution and no squared terms (first-order model without interaction). An increase in the injection time resulted in a decrease of the resolution (Fig. 1b). The regular curvature in the response surface indicated a second-order model without interaction.

Appropriate software indicates the region in which the response will meet the method requirements and if several responses are to be satisfied, e.g. resolution between several pairs of compounds, the response contours can be overlaid so that the best compromise can be chosen.

5. Conclusion

This paper has attempted to cover the main aspects of robustness testing in LC and CE and to outline the information it provides from some experimental designs which have been used. It has been shown that screening designs may be sufficient to set the method limits, which is an essential prerequisite to ensure the reliability of its results in routine use. Calculations can be performed manually or with classical statistical software packages. Design softwares (e.g. Design Ease, Nemrod, etc.) are also valuable tools in the choice of fractional factorial designs. Response surface methodology requires special computer packages (e.g. SAS, RS discover, Design Expert, Nemrod, etc.) for the design selection, determination of the region corresponding to the best compromise for the responses and response prediction. Response surfaces are of major interest in method transfer because they give a comprehensive picture of the behavior and limitations of the method. The information provided by robustness testing shows that it is an integral part of method validation. It is recommended to include the results of robustness testing in the registration dossier.

Acknowledgements

S.D. Filbey and Dr. D.D. Rudd, from Glaxo Research and Development, Ware, UK, and Dr. E. Puech from the University Paul Sabatier, Toulouse, France, are thanked for helpful comments and discussions.

References

[1] Validation of Analytical Procedures, Guidelines prepared within the International Conference on Harmonisation, Commision of the European Communities, Brussels, 1995, ref. III/5626/94.

[2] USP XXIII, 1982–1984, Mack Publishing Company, Easton, PA, 1995.

[3] Explanatory Note, Analytical Validation, Office for Official Publications of the European Communities, Luxembourg, 75/318/CEE modified August 1989, ref. III/844/87.

[4] H. Fabre, V. Meynier de Salinelles and B. Mandrou, Analusis, 118 (1985) 1061–1064.

[5] H. Fabre, M. Sekkat, M.D. Blanchin and B. Mandrou, J. Pharm. Biomed. Anal., 7 (1989) 1711–1718.

[6] S. Sun, H. Fabre and H. Maillols, J. Liq. Chromatogr., 17 (1994) 2495–2509.

[7] M. Mulholland and J. Waterhouse, J. Chromatogr., 395 (1987) 539–551.

[8] M. Mulholland, Trends Anal. Chem., 7 (1988) 383–389.

[9] M. Mulholland, and J. Waterhouse, Chromatographia, 25 (1988) 769–774.

[10] J.A. Van Leeuwen, B.G.M. Vandeginste, G. Kateman, M. Mulholland and A. Cleland, Anal. Chim. Acta, 228 (1990) 145–153.

[11] J.A. Van Leeuwen, L.M.C. Buydens, B.G.M. Vandeginste, G. Kateman, P.J. Schenmakers and M. Mulholland, Chemomet. Intell. Lab. Syst., 11 (1991) 37–55.

[12] Y. Vander Heyden, Analusis, 22 (1994) M27–M29.

[13] Guide de Validation Analytique, Rapport d'une commission de la SFSTP, STP Pharm. Prat., 2 (1992) 205–226.

[14] P. Chaminade, S. Feraud, A. Baillet and D. Ferrier, STP Pharm. Prat., 5 (1995) 17–35.

[15] J.L. Virlichie and A. Ayache, STP Pharm. Prat., 5 (1995) 49–60.

[16] K.D. Altria and S.D. Filbey, Chromatographia, 39 (1994) 306–310.

[17] S.D. Filbey and K.D. Altria, J. Capillary E.ectrophoresis, 1 (1994) 90.

[18] G.E.P. Box, W.G. Hunter and J.S. Hunter, Statistics for Experimenters. An Introduction to Design, Data Analysis and Model Building, Wiley, New York, 1978, Part III, pp. 291–453.

[19] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, Chemometrics: A Textbook, Elsevier, Amsterdam, 1988, pp. 101–106, 271–291.

[20] S.N. Deming and S.L. Morgan, Experimental Design: A Chemometric Approach, Elsevier, Amsterdam, 1987.

[21] E. Morgan, Chemometrics: Experimental Design, Wiley, New York, 1991.

[22] O.L. Davies and P.L. Goldsmith (Eds.), Statistical Methods in Research and Production, 4th revised edn., Longmans, London, 1978.

[23] L. Davies, Efficieny in Research, Development, and Production: The Statistical Design and Analysis of Chemical Experiments, Royal Society of Chemistry, Cambridge, 1993.

[24] D.J. Wheeler, Tables of Screening Designs, 2nd edn., SPC Press, Knoxville, TX, 1989.

[25] G.T. Wernimont, Use of Statistics to Develop and Evaluate Analytical Methods, The Association of Official Analytical Chemists, Arlington, VA, 1985.

[26] International Organization for Standardization, Accuracy (Trueness and Precision) of Measurements, Methods and Results, ISO/DIS 5725-1 to 5725-3, draft versions 1990/91.

[27] W.J. Youden and E.H. Steiner, Statistical Manual of the Association of Official Analytical Chemists, The Association of Official Analytical Chemists, Washington DC, 1975 pp. 33–41.

[28] S.D. Filbey, personal communication, 1994.