Paul De Bièvre · Helmut Günzler (Eds.)

Validation in Chemical Measurement

Springer

Prof. Dr. Paul De Bièvre
Duineneind 9
2460 Kasterlee
Belgium

Prof. Dr. Helmut Günzler
Bismarckstrasse 11
69469 Weinheim
Germany

Library of Congress Control Number: 2004114858

ISBN 3-540-20788-0 **Springer Berlin Heidelberg New York**

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springeronline.com
© Springer-Verlag Berlin Heidelberg 2005
Printed in Germany

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: *design & production* GmbH, Heidelberg
Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig

Printed on acid-free paper 52/3111/YL - 5 4 3 2 1 0

Preface

Validation of measurement methods has been used for a very long time in chemistry. It is mostly based on the examination of a measurement procedure for its characteristics such as precision, accuracy, selectivity, sensitivity, repeatability, reproducibility, detection limit, quantification limit and more.

When focussing on quality comparability and reliability in chemical measurement, the fields of interest to this Journal, one stumbles into various interpretations of the term *validation*. It is one more example of a term which is used sometimes very consistently, sometimes very loosely or indeed ambiguously. Since the term is very common in the chemical community, it is important that its meaning be clear. Turning to the 2nd edition of the International Vocabulary of Basic and General terms in Metrology (VIM) (1993), surprisingly we do not find a definition. Webster's Dictionary of the English language (1992) tells us that validation is 'making or being made valid'. Obviously *validation* has to do with *valid.* The same Webster indicates the meaning of the corresponding verb: to validate seems 'to make valid or binding, to confirm the validity of (Latin: validare)', where *valid* means: 'seen to be in agreement with the facts or to be logically sound'. We certainly can build on this to have a 'valid' discussion. Validation of a method clearly seems to mean making 'valid' the measurement results obtained by this method. The first definition 'seen to be in agreement with the facts', is rather difficult to apply. The second definition however, tells us that 'validation of a method is to make the method to be seen as logically sound'. It looks as if validation of a method is a process whereby it is tested and demonstrated by somebody or some authority to be logically sound. Such a validation should enable everybody to use it. That implies a list of methods 'validated' by competent authorities in the field concerned, which sounds possible and useful. Is that not what AOAC does?

Sometimes, the notion of validating a measurement result also shows up. Apparently it means to make a result 'valid', and even binding, i.e. confirming its 'validity'. Since *valid* means 'seen to be in agreement with the facts', that almost sounds as a synonym for 'accurate'. That makes sense and there seems to be no argument as to whether a method or a result can be validated (they can). An important question arises: does a validated method automatically give a validated measurement result, i.e. a quantity value[1] with associated measurement uncertainty? The answer must be: no. There can never be a mechanism or recipe for producing *automatically* 'valid' results because one can never eliminate the skills, the role and the responsibility of the analyst.

ISO 9000:2000, item 3.8.5 defines *validation* as 'confirmation by examination and provision of objective evidence that the requirements for an intended use are fulfilled'. The revised edition of the VIM ('VIM3'), is likely to fine-tune this definition of the concept 'validation' to be 'confirmation through examination of a given item and provision of objective evidence that it fulfills the requirements for a stated intended use'.

Looking at simple practice, many people are looking for a formal decision that a given measurement method *automatically* gives them '*valid*' i.e. reliable results. One wonders what this has to do with 'stated intended use'. Reliability clearly is a property of a measurement result. Checking whether that result fulfills the requirement for a stated intended use, seems to be a totally different matter. That requires the formulation of a requirement a priori, i.e. before the measurement is made, and derived from the need for a measurement result, not from the result itself.

This anthology contains 31 outstanding papers published in the Journal "Accreditation and Quality Assurance" since its inception, but mostly in the period 2000–2003, on the topic 'validation'. They reflect the latest understanding – or lack thereof –, of the concept and possibly some rationale(s) for the answer to the question why it is important to integrate the concept of 'validation' into the standard procedures of every measurement laboratory.

It is hoped that this anthology is of benefit to both the producers and the users of results of chemical measurements: the basic concepts and the basic thinking in measurement are the same for both.

Prof. Dr. P. De Bièvre
Editor-in-Chief
Accreditation and Quality Assurance
Kasterlee 2004-04-02

[1] quantity (German: 'Messgrösse', French: 'grandeur de mesure', Dutch: 'meetgrootheid') is not used here in the meaning 'amount', but as the generic term for the quantities we measure: concentration, volume, mass, temperature, time, etc., as defined in the VIM.

Contents

Contributors

Hassan Y. Aboul-Enein
Biological and Medical Research Department (MBC-03), King Faisal Specialist Hospital & Research Centre, P.O. Box 3354, Riyadh 11211, Saudi Arabia

G. Anand Kumar
VIMTA Labs Ltd., Hyderabad–500 051, India

Boris Anisimov
The National Physical Laboratory of Israel, Dancinger A. Bldg., Givat Ram, Jerusalem 91904, Israel

Elke Anklam
European Commission, Directorate General Joint Research Centre, Institute for Reference Materials and Measurements, Food Safety and Quality Unit, Retieseweg 111, B-2440 Geel, Belgium

Vicki J. Barwick
Laboratory of the Government Chemist, Queens Road, Teddington, Middlesex, TW11 0LY, UK

Ricardo J. N. Bettencourt da Silva
CECUL, Faculdade de Ciências da Universidade de Lisboa, P-1700 Lisbon, Portugal

M. Buzoianu
National Institute of Metrology, Sos. Vitan-Bârzesti 11, 75669 Bucharest, Romania

M. Filomena G. F. C. Camões
CECUL, Faculdade de Ciências da Universidade de Lisboa, P-1700 Lisbon, Portugal

J. J. Cuello
University of Zaragoza, Analytical Chemistry Department. Crtra. Zaragoza, s/n 22071 Huesca, Spain

R. C. Díaz
University of Zaragoza, Analytical Chemistry Department. Crtra. Zaragoza, s/n 22071 Huesca, Spain

Stephen L. R. Ellison
Laboratory of the Government Chemist, Queens Road, Teddington, Middlesex, TW11 0LY, UK

Rattanjit S. Gill
Laboratory of the Government Chemist, Queens Road, Teddington, Middlesex, TW11 0LY, UK

Alison Gillespie
Laboratory of the Government Chemist, Queens Road, Teddington, Middlesex, TW11 0LY, UK

Mykolas Gladkovas
Institute of Chemistry, Goštauto, 92600 Vilnius, Lithuania

H. Hey
Official food and veterinary laboratory of Schleswig-Holstein, 24537 Neumünster, Germany

D. Brynn Hibbert
Department of Analytical Chemistry, University of New South Wales, Sydney, NSW 2052, Australia

Magnus Holmgren
SP Swedish National Testing and Research Institute, P.O. Box 857, SE-50115 Borås, Sweden

Ludwig Huber
Hewlett-Packard GmbH Waldbronn, P.O. Box 1280, 76337 Waldbronn, Germany

Rouvim Kadis
D. I. Mendeleyev Institute for Metrology, 19 Moskovsky pr., 198005 St. Petersburg, Russia

H. Kohno
Quality Assurance Department, Sandoz Pharmaceuticals Ltd., No. 1 Takeno, Kawagoe City, Saitama, Japan

Reinhard Kreßner
Federal Environmental Agency, P. O. Box 33 00 22, 14191 Berlin, Germany

Stephan Küppers
Schering AG, In-Process-Control, Müllerstrasse 170–178, 13342 Berlin, Germany

Ilya Kuselman
The National Physical Laboratory of Israel, Dancinger A. Bldg., Givat Ram, Jerusalem 91904, Israel

Margreet Lauwaars
European Commission, Directorate General Joint Research Centre, Institute for Reference Materials and Measurements, Food Safety and Quality Unit, Retieseweg 111, B-2440 Geel, Belgium

Alexandre Leclercq
Institut Pasteur de Lille, Department of Food Microbiology and Safety, Rue du Professeur Calmette 1, 59000 Lille, France

Alex Lepek
Tech Projects, P.O. Box 9769, Jerusalem 91091, Israel

Peter Lepom
Federal Environmental Agency, P. O. Box 33 00 22, 14191 Berlin, Germany

Solveig Linko
Medix Diacor Labservice, Department of Quality Assurance, Nihtisillankuja 1, 00260 Espoo, Finland

Nineta Majcen
Metrology Institute of the Republic of Slovenia (MIRS), Laboratory Centre Celje, Tkalska 15, SI-3000 Celje, Slovenia

Hans H. Maurer
Department of Experimental and Clinical Toxicology, Institute of Experimental and Clinical Pharmacology and Toxicology, University of Saarland, 66421 Homburg, Germany

S. Othake
Quality Assurance Department, Sandoz Pharmaceuticals Ltd., No. 1 Takeno, Kawagoe City, Saitama, Japan

Frank T. Peters
Department of Experimental and Clinical Toxicology, Institute of Experimental and Clinical Pharmacology and Toxicology, University of Saarland, 66421 Homburg, Germany

Mark J. Q. Rafferty
Laboratory of the Government Chemist, Queens Road, Teddington, Middlesex, TW11 0LY, UK

Eugenija Ramoškienė
Institute of Chemistry, Goštauto, 92600 Vilnius, Lithuania

C. Ríos
University of Zaragoza, Analytical Chemistry Department. Crtra. Zaragoza, s/n 22071 Huesca, Spain

Gordon A. Ross
Hewlett-Packard GmbH Waldbronn, 76337 Waldbronn, Germany

H. S. Sahoo
VIMTA Labs Ltd., Hyderabad–501 051, India

Mudis Šalkauskas
Institute of Chemistry, Goštauto, 92600 Vilnius, Lithuania

M. A. Sarasa
University of Zaragoza, Analytical Chemistry Department. Crtra. Zaragoza, s/n 22071 Huesca, Spain

S. K. Satpathy
VIMTA Labs Ltd., Hyderabad–502 051, India

João Seabra e Barros
Instituto Nacional de Engenharia e Tecnologia Industrial, Estrada do Paço do Lumiar, P-1699 Lisbon Codex, Portugal

S. Seno
Quality Assurance Department, Sandoz Pharmaceuticals Ltd., No. 1 Takeno, Kawagoe City, Saitama, Japan

Avinoam Shenhar
The National Physical Laboratory of Israel, Dancinger A. Bldg., Givat Ram, Jerusalem 91904, Israel

Raluca-Iona Stefan
Department of Analytical Chemistry, Faculty of Chemistry, University of Bucharest, Blvd. Republicii No. 13, 70346 Bucharest-3, Romania

G. Swarnabala
VIMTA Labs Ltd., Hyderabad–503 051, India

Sue Upton
Laboratory of the Government Chemist, Queens Road, Teddington, Middlesex, TW11 0LY, UK

Robert J. Wells
Australien Government Analytical Laboratories, 1 Suakin Street, Pymble, NSW 2073, Australia

Herbert Wiederoder
Hewlett-Packard GmbH Waldbronn, P. O. Box 1280, 76337 Waldbronn, Germany

Alex Williams
19 Hamesmoor Way, Mytchett, Camberley, Surrey, GU16 6JG, UK

Peter Woitke
Federal Environmental Agency, P. O. Box 33 00 22, 14191 Berlin, Germany

Accred Qual Assur (2002) 7:441–449
DOI 10.1007/s00769-002-0516-5

© Springer-Verlag 2002

Frank T. Peters
Hans H. Maurer

Bioanalytical method validation and its implications for forensic and clinical toxicology – A review

Received: 16 June 2002
Accepted: 12 July 2002

Part of this review was published in the communications of the International Association of Forensic Toxicologists (TIAFT; TIAFT Bulletin 32 (2002): 16–23) and of the Society for Forensic and Toxicologic Chemistry (GTFCH; Toxichem and Krimitech 68 (2001): 116-126).

F.T. Peters (✉) · H.H. Maurer
Department of Experimental
and Clinical Toxicology,
Institute of Experimental and Clinical
Pharmacology and Toxicology,
University of Saarland, 66421 Homburg,
Germany
e-mail: frank.peters@uniklinik-saarland.de
Tel.: +49-6841-1626050
Fax: +49-6841-1626051

Abstract The reliability of analytical data is very important to forensic and clinical toxicologists for the correct interpretation of toxicological findings. This makes (bio)analytical method validation an integral part of quality management and accreditation in analytical toxicology. Therefore, consensus should be reached in this field on the kind and extent of validation experiments as well as on acceptance criteria for validation parameters. In this review, the most important papers published on this topic since 1991 have been reviewed. Terminology, theoretical and practical aspects as well as implications for forensic and clinical toxicology of the following validation parameters are discussed: selectivity (specificity), calibration model (linearity), accuracy, precision, limits, stability, recovery and ruggedness (robustness).

Keywords Method · Validation · Bioanalysis · Toxicology

Introduction

The reliability of analytical findings is a matter of great importance in forensic and clinical toxicology, as it is a prerequisite for correct interpretation of toxicological findings. Unreliable results might not only be contested in court, but could also lead to unjustified legal consequences for the defendant or to wrong treatment of the patient. The importance of validation, at least of routine analytical methods, can therefore hardly be overestimated. This is especially true in the context of quality management and accreditation, which have become matters of increasing importance in analytical toxicology in recent years. This is also reflected in the increasing requirements of peer-reviewed scientific journals concerning method validation. Therefore, this topic should be extensively discussed on an international level to reach a consensus on the extent of validation experiments and on acceptance criteria for validation parameters of bioanalytical methods in forensic and clinical toxicology.

Over the last decade, similar discussions have been going on in the closely related field of pharmacokinetic studies for registration of pharmaceuticals. This is reflected by the number of publications on this topic published in the last decade, of which the most important are discussed here.

Important publications on validation (1991 to present)

A review on validation of bioanalytical methods was published by Karnes et al. in 1991 which was intended to provide guidance for bioanalytical chemists [1]. One

year later, Shah et al. published their report on the conference on "Analytical Methods Validation: Bioavailability, Bioequivalence and Pharmacokinetic Studies" held in Washington in 1990 (Conference Report) [2]. During this conference, consensus was reached on which parameters of bioanalytical methods should be evaluated, and some acceptance criteria were established. In the following years, this report was actually used as guidance by bioanalysts. Despite the fact, however, that some principle questions had been answered during this conference, no specific recommendations on practical issues like experimental designs or statistical evaluation were made. In 1994, Hartmann et al. analysed the Conference Report performing statistical experiments on the established acceptance criteria for accuracy and precision [3]. Based on their results they questioned the suitability of these criteria for practical application. From 1995 to 1997, application issues like experimental designs and statistical methods for bioanalytical method validation were discussed in a number of publications by Dadgar et al. [4, 5], Wieling et al. [6], Bressolle et al. [7] and Causon [8]. An excellent review on validation of bioanalytical chromatographic methods was published by Hartmann et al. in 1998, in which theoretical and practical issues were discussed in detail [9]. In an update of the Washington Conference in 2000, experiences and progress since the first conference were discussed. The results were again published by Shah et al. in a report (Conference Report II) [10], which has also been used as a template for guidelines drawn up by the U.S. Food and Drug Administration (FDA) for their own use [11]. Besides, it should be mentioned that some journals like the Journal of Chromatography B [12] or Clinical Chemistry have established their own criteria for validation. Two other documents that seem to be important in this context have been developed by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) and approved by the regulatory agencies of the European Union, the United States of America and Japan. The first, approved in 1994, concentrated on the theoretical background and definitions [13], the second, approved in 1996, on methodology and practical issues [14]. Both can be downloaded from the ICH homepage free of charge (www.ich.org). Finally, in 2001 Vander Heyden et al. published a paper on experimental designs and evaluation of robustness/ruggedness tests [15]. Despite the fact that the three last mentioned publications were not especially focussed on bioanalytical methods, they still contain helpful guidance on some principal questions and definitions in the field of analytical method validation.

The aim of our review is to present and compare the contents of the above mentioned publications on (bio)analytical method validation, and to discuss possible implications for forensic and clinical toxicology.

Terminology

The first problem encountered when studying literature on method validation are the different sets of terminology employed by different authors. A detailed discussion of this problem can be found in the review of Hartmann et al. [9]. Therein, it was proposed to adhere, in principle, to the terminology established by the ICH [13], except for accuracy, for which the use of a more detailed definition was recommended (cf. Accuracy). However, the ICH terminology lacked a definition for stability, which is an important parameter in bioanalytical method validation. Furthermore, the ICH definition of selectivity did not take into account interferences that might occur in bioanalysis (e.g. from metabolites). For both parameters, however, reasonable definitions were provided by Conference Report II [10].

Validation parameters

There is a general agreement that at least the following validation parameters should be evaluated for quantitative procedures: selectivity, calibration model (linearity), stability, accuracy (bias, precision) and limit of quantification. Additional parameters which might have to be evaluated include limit of detection, recovery, reproducibility and ruggedness (robustness) [2, 4–10, 12].

Selectivity (specificity)

In Conference Report II, selectivity was defined as follows: "Selectivity is the ability of the bioanalytical method to measure unequivocally and to differentiate the analyte(s) in the presence of components, which may be expected to be present". Typically, these might include metabolites, impurities, degradants, matrix components, etc. [10]. This definition is very similar to the one established by the ICH [13], but takes into account the possible presence of metabolites, and thus is more applicable for bioanalytical methods.

There are two points of view on when a method should be regarded as selective. One way to establish method selectivity is to prove the lack of response in blank matrix [1, 2, 4–10, 12, 14]. The requirement established by the Conference Report [2] to analyse at least six different sources of blank matrix has become state of the art. However, this approach has been subject to criticism in the review of Hartmann et al., who stated from statistical considerations, that relatively rare interferences will remain undetected with a rather high probability [9]. For the same reason, Dadgar et al. proposed to evaluate at least 10–20 sources of blank samples [4]. However, in Conference Report II [10], even analysis of only one source of blank matrix was deemed acceptable, if

hyphenated mass spectrometric methods are used for detection.

The second approach is based on the assumption that small interferences can be accepted as long as precision and bias remain within certain acceptance limits. This approach was preferred by Dadgar et al. [4] and Hartmann et al. [9]. Both publications proposed analysis of up to 20 blank samples spiked with analyte at the lower limit of quantification (LLOQ) and, if possible, with interferents at their highest likely concentrations. In this approach, the method can be considered sufficiently selective if precision and accuracy data for these LLOQ samples are acceptable. For a detailed account of experimental designs and statistical methods to establish selectivity see Ref. [4].

Whereas the selectivity experiments for the first approach can be performed during a pre-validation phase (no need for quantification), those for the second approach are usually performed together with the precision and accuracy experiments during the main validation phase.

At this point it must be mentioned that the term specificity is used interchangeably with selectivity, although in a strict sense specificity refers to methods which produce a response for a single analyte, whereas selectivity refers to methods that produce responses for a number of chemical entities, which may or may not be distinguished [1]. Selective multi-analyte methods (e.g. for different drugs of abuse in blood) should of course be able to differentiate all interesting analytes from each other and from the matrix.

Calibration model (linearity)

The choice of an appropriate calibration model is necessary for reliable quantification. Therefore, the relationship between the concentration of analyte in the sample and the corresponding detector response must be investigated. This can be done by analysing spiked calibration samples and plotting the resulting responses versus the corresponding concentrations. The resulting standard curves can then be further evaluated by graphical or mathematical methods, the latter also allowing statistical evaluation of the response functions.

Whereas there is general agreement that calibration samples should be prepared in blank matrix and that their concentrations must cover the whole calibration range, recommendations on how many concentration levels should be studied with how many replicates per concentration level differ significantly [5–10, 12]. In Conference Report II, a sufficient number of standards to define adequately the relationship between concentration and response was demanded. Furthermore, it was stated that at least five to eight concentration levels should be studied for linear and maybe more for non-linear relationships [10]. However, no information was given on how many replicates should be analysed at each level. The guidelines established by the ICH and those of the Journal of Chromatography B also required at least five concentration levels, but again no specific requirements for the number of replicates at each level were given [12, 14]. Causon recommended six replicates at each of six concentration levels, whereas Wieling et al. used eight concentration levels in triplicate [6, 8]. Based on studies by Penninckx et al. [16], Hartmann et al. proposed in their review to rather use fewer concentration levels with a greater number of replicates (e.g. four evenly spread levels with nine replicates) [9]. This approach not only allows the reliable detection of outliers, but also a better evaluation of the behaviour of variance across the calibration range. The latter is important for choosing the right statistical model for the evaluation of the calibration curve. The often used ordinary least squares model for linear regression is only applicable for homoscedastic data sets (constant variance over the whole range), whereas in case of heteroscedasticity (significant difference between variances at lowest and highest concentration levels) the data should mathematically be transformed or a weighted least squares model should be applied [6–10]. Usually, linear models are preferable but, if necessary, the use of non-linear models is not only acceptable but even recommended. However, more concentration levels are needed for the evaluation of non-linear models than for linear models [2, 9, 10].

After outliers have been purged from the data and a model has been evaluated visually and/or by, e.g. residual plots, the model fit should also be tested by appropriate statistical methods [2, 6, 9, 10, 14]. The fit of unweighted regression models (homoscedastic data) can be tested by the ANOVA lack-of-fit test [6, 9]. A detailed discussion of alternative statistical tests for both unweighted and weighted calibration models can be found in Ref. [16]. The widespread practice to evaluate a calibration model via its coefficients of correlation or determination is not acceptable from a statistical point of view [9].

However, one important point should be kept in mind when statistically testing the model fit: The higher the precision of a method, the higher the probability to detect a statistically significant deviation from the assumed calibration model [1, 6, 9]. Therefore, the practical relevance of the deviation from the assumed model should also be taken into account. If the accuracy data (bias and precision) are within the required acceptance limits or an alternative calibration model is not applicable, slight deviations from the assumed model may be neglected [6, 9].

Once a calibration model has been established, the calibration curves for other validation experiments (precision, bias, stability, etc.) and for routine analysis can be prepared with fewer concentration levels and fewer or no replicates [6, 9].

Accuracy

The accuracy of a method is affected by systematic (bias) as well as random (precision) error components [3, 9] This fact has been taken into account in the definition of accuracy as established by the International Organization for Standardization (ISO) [17]. However, it must be mentioned that accuracy is often used to describe only the systematic error component, i.e. in the sense of bias [1, 2, 6–8, 10, 12, 13]. In the following, the term accuracy will be used in the sense of bias, which will be indicated in brackets.

Bias

According to ISO, bias is the difference between the expectation of the test results and an accepted reference value [17]. It may consist of more than one systematic error component. Bias can be measured as a percent deviation from the accepted reference value. The term trueness expresses the deviation of the mean value of a large series of measurements from the accepted reference value. It can be expressed in terms of bias.

Due to the high workload of analysing such large series, trueness is usually not determined during method validation, but rather from the results of a great number of quality control (QC) samples during routine application or in interlaboratory studies.

Precision

According to ICH, precision is the closeness of agreement (degree of scatter) between a series of measurements obtained from multiple sampling of the same homogenous sample under the prescribed conditions and may be considered at three levels: repeatability, intermediate precision and reproducibility [13]. Precision is usually measured in terms of imprecision expressed as an absolute or relative standard deviation and does not relate to reference values.

Repeatability. Repeatability expresses the precision under the same operating conditions over a short interval of time. Repeatability is also termed intra-assay precision [13]. Within-run or within-day precision are also often used to describe repeatability.

Intermediate precision. Intermediate precision expresses within-laboratories variations: different days, different analysts, different equipment, etc. [13]. The ISO definition used the term "M-factor different intermediate precision", where the M-factor expresses how many and which factors (time, calibration, operator, equipment or combinations of those) differ between successive determinations [17]. In a strict sense, intermediate precision is the total precision under varied conditions, whereas so called inter-assay, between-run or between-day precision only measure the precision components caused by the respective factors. However, the latter terms are not clearly defined and obviously often used interchangeably with each other and also with the term intermediate precision.

Reproducibility. Reproducibility expresses the precision between laboratories (collaborative studies, usually applied to standardization of methodology) [13]. Reproducibility only has to be studied, if a method is supposed to be used in different laboratories.

Unfortunately, some authors also used the term reproducibility for within-laboratory studies at the level of intermediate precision [8, 12]. This should, however, be avoided in order to prevent confusion.

As already mentioned above, precision and bias can be estimated from the analysis of QC samples under specified conditions. As both precision and bias can vary substantially over the calibration range, it is necessary to evaluate these parameters at least at three concentration levels (low, medium and high relative to the calibration range) [1, 2, 9, 10, 14]. In Conference Report II, it was further defined that the concentration of the low QC sample must be within three times LLOQ [10]. The Journal of Chromatography B requirement is to study precision and bias at two concentration levels (low and high), whereas in the experimental design proposed by Wieling et al. four concentration levels (LLOQ, low, medium, high) were studied [6, 12]. Causon also suggested to estimate precision at four concentration levels [8]. Several authors have specified acceptance limits for precision and/or accuracy (bias) [2, 7, 8, 10, 12]. Both Conference Reports required precision to be within 15% relative standard deviation (RSD) except at the LLOQ where 20% RSD was accepted. Bias was required to be within ±15% of the accepted true value, except at the LLOQ where ±20% were accepted [2, 10]. These requirements have been subject to criticism in the analysis of the Conference Report by Hartmann et al. [3]. They concluded from statistical considerations that it is not realistic to apply the same acceptance criteria at different levels of precision (repeatability, reproducibility) as RSD under reproducibility conditions is usually considerably greater than under repeatability conditions. Furthermore, if precision and bias estimates are close to the acceptance limits, the probability to reject an actually acceptable method (β-error) is quite high. Causon proposed the same acceptance limits of 15% RSD for precision and ±15% for accuracy (bias) for all concentration levels [8].

The guidelines established by the Journal of Chromatography B required precision to be within 10% RSD for the high QC samples and within 20% RSD for the low QC sample. Acceptance criteria for accuracy (bias) were not specified therein [12].

Again, the proposals on how many replicates at each concentration levels should be analysed vary considerably. The Conference Reports and Journal of Chromatography B guidelines required at least five replicates at each concentration level [2, 10, 12]. However, one would assume that these requirements only apply to repeatability studies; at least no specific recommendations were given for studies of intermediate precision or reproducibility. Some more practical approaches to this problem have been described by Wieling et al. [6], Causon [8] and Hartmann et al. [9]. In their experimental design, Wieling et al. analysed three replicates at each of four concentration levels on each of 5 days. Similar approaches were suggested by Causon (six replicates at each of four concentrations on each of four occasions) and Hartmann et al. (two replicates at each concentration level on each of 8 days). All three used or proposed one-way ANOVA to estimate repeatability and time-different precision components. In the design proposed by Hartmann et al. the degrees of freedom for both estimations are most balanced, namely 8 for within-run precision and 7 for between-run precision. In the information for authors of the Clinical Chemistry journal, an experimental design with two replicates per run, two runs per day over 20 days for each concentration level is recommended, which has been established by the NCCLS [18]. This not only allows estimation of within-run and between-run standard deviations, but also of within-day, between-day and total standard deviations, which are in fact all estimations of precision at different levels. However, it seems questionable if the additional information provided by this approach can justify the high workload and costs compared to the other experimental designs.

Daily variations of the calibration curve can influence bias estimation. Therefore, bias estimation should be based on data calculated from several calibration curves [9]. In the experimental design of Wieling et al., the results for QC samples were calculated via daily calibration curves. Therefore, the overall means from these results at the different concentration levels reliably reflect the average bias of the method at the corresponding concentration level. Alternatively, as described in the same paper, the bias can be estimated using confidence limits around the calculated mean values at each concentration [6]. If the calculated confidence interval includes the accepted true value, one can assume the method to be free of bias at a given level of statistical significance. Another way to test the significance of the calculated bias is to perform a *t*-test against the accepted true value.

However, even methods exhibiting a statistically significant bias can still be acceptable, if the calculated bias lies within previously established acceptance limits. Other methods for bias evaluation can be found in Ref. [9].

Limits

Lower limit of quantification (LLOQ)

The LLOQ is the lowest amount of an analyte in a sample that can be quantitatively determined with suitable precision and accuracy (bias) [10, 13]. There are different approaches for the determination of LLOQ.

LLOQ based on precision and accuracy (bias) data [2, 7–10, 13, 14]. This is probably the most practical approach and defines the LLOQ as the lowest concentration of a sample that can still be quantified with acceptable precision and accuracy (bias). In the Conference Reports, the acceptance criteria for these two parameters at LLOQ are 20% RSD for precision and ±20% for bias. Only Causon suggested 15% RSD and ±15%, respectively [8]. It should be pointed out, however, that these parameters must be determined using an LLOQ sample independent from the calibration curve. The advantage of this approach is the fact that the estimation of LLOQ is based on the same quantification procedure used for real samples.

LLOQ based on signal to noise ratio (S/N) [12, 14]. This approach can only be applied if there is baseline noise, e.g. to chromatographic methods. Signal and noise can then be defined as the height of the analyte peak (signal) and the amplitude between the highest and lowest point of the baseline (noise) in a certain area around the analyte peak. For LLOQ, S/N is usually required to be equal to or greater than 10.

The estimation of baseline noise can be quite difficult for bioanalytical methods, if matrix peaks elute close to the analyte peak.

LLOQ based on standard deviation of the response from blank samples [14]. Another definition of LLOQ is the concentration that corresponds to a detector response that is k times greater than the estimated standard deviation of blank samples (SD_{bl}). From the detector signal, the LLOQ can be calculated using the slope of the calibration curve (S) with following formula: $LLOQ = k \cdot SD_{bl}/S$ (for blank corrected signals).

This approach is only applicable for methods where SD_{bl} can be estimated from replicate analysis of blank samples. It is therefore not applicable for most quantitative chromatographic methods, as here the response is usually measured in terms of peak area units, which can of course not be measured in a blank sample analysed with a selective method.

LLOQ based on a specific calibration curve in the range of LLOQ [14]. In this approach, a specific calibration curve is established from samples containing the analyte in the range of LLOQ. One must not use the calibration

curve over the whole range of quantification for this determination. The standard deviation of the blank can then be estimated from the residual standard deviation of the regression line or the standard deviation of the y intercept. The calculations of LLOQ are basically the same as described under the heading "LLOQ based on standard deviation of the response from the blank samples". This approach is also applicable for chromatographic methods.

Upper limit of quantification (ULOQ)

The upper limit of quantification is the maximum analyte concentration of a sample that can be quantified with acceptable precision and accuracy (bias). In general the ULOQ is identical to the concentration of the highest calibration standard [10].

Limit of detection (LOD)

Quantification below LLOQ is by definition not acceptable [2, 5, 9, 10, 13, 14]. Therefore, below this value a method can only produce semiquantitative or qualitative data. However, it can still be important to know the LOD of the method. According to ICH, it is the lowest concentration of analyte in a sample which can be detected but not necessarily quantified as an exact value. According to Conference Report II, it is the lowest concentration of an analyte in a sample, that the bioanalytical procedure can reliably differentiate from background noise [10, 13].

The approaches for estimation of the LOD are basically the same as those described for LLOQ under the headings "LLOQ based on signal to noise ratio (S/N)" – "LOQ based on a specific calibration curve in the range of LLOQ". However, for LOD a S/N or k-factor equal to or greater than three is usually chosen [1, 6, 9, 14]. If the calibration curve approach is used for determination of the LOD, only calibrators containing the analyte in the range of LOD must be used.

Stability

The definition according to Conference Report II was as follows: "The chemical stability of an analyte in a given matrix under specific conditions for given time intervals" [10]. Stability of the analyte during the whole analytical procedure is a prerequisite for reliable quantification. Therefore, full validation of a method must include stability experiments for the various stages of analysis including storage prior to analysis.

Long-term stability

The stability in the sample matrix should be established under storage conditions, i.e. in the same vessels, at the same temperature and over a period at least as long as the one expected for authentic samples [1, 2, 4, 5, 9, 10, 12].

Freeze/thaw stability

As samples are often frozen and thawed, e.g. for re-analysis, the stability of analyte during several freeze/thaw cycles should also be evaluated. The Conference Reports require a minimum of three cycles at two concentrations in triplicate, which has also been accepted by other authors [2, 4, 6, 9, 10].

In-process stability

The stability of analyte under the conditions of sample preparation (e.g. ambient temperature over time needed for sample preparation) is evaluated here. There is general agreement, that this type of stability should be evaluated to find out, if preservatives have to be added to prevent degradation of analyte during sample preparation [4, 9, 10].

Processed sample stability

Instability cannot only occur in the sample matrix, but also in prepared samples. It is therefore important to also test the stability of an analyte in the prepared samples under conditions of analysis (e.g. autosampler conditions for the expected maximum time of an analytical run). One should also test the stability in prepared samples under storage conditions, e.g. refrigerator, in case prepared samples have to be stored prior to analysis [4–6, 9, 10].

For more details on experimental design and statistical evaluation of stability experiments see Refs. [4, 5, 9].

Stability can be tested by comparing the results of QC samples analysed before (comparison samples) and after (stability samples) being exposed to the conditions for stability assessment. It has been recommended to perform stability experiments at least at two concentration levels (low and high) [4–6, 9]. For both, comparison and stability samples, analysis of at least six replicates was recommended [9]. Ratios between comparison samples and stability samples of 90–110% with 90% confidence intervals within 80–120% [9] or 85–115% [4] were regarded acceptable. Alternatively, the mean of the reference samples can be tested against a lower acceptance limit corresponding to 90% of the mean of the comparison samples [8, 9].

Recovery

As already mentioned above, recovery is not among the validation parameters regarded as essential by the Conference Reports. Most authors agree, that the value for recovery is not important, as long as the data for LLOQ, (LOD), precision and accuracy (bias) are acceptable [1, 5, 7–10]. It can be calculated as the percentage of the analyte response after sample workup compared to that of a solution containing the analyte at the theoretical maximum concentration. Therefore, absolute recoveries can usually not be determined if the sample workup includes a derivatization step, as the derivatives are often not available as reference substances. Nevertheless, the guidelines of the Journal of Chromatography B require the determination of the recovery for analyte and internal standard at high and low concentrations [12].

Ruggedness (robustness)

Ruggedness is a measure for the susceptibility of a method to small changes that might occur during routine analysis like small changes of pH values, mobile phase composition, temperature, etc. Full validation must not necessarily include ruggedness testing; it can however be very helpful during the method development/pre-validation phase, as problems that may occur during validation are often detected in advance. Ruggedness should be tested, if a method is supposed to be transferred to another laboratory [1, 9, 13–15]. A detailed account and helpful guidance on experimental designs and evaluation of ruggedness/robustness tests can be found in Ref. [15].

Implications for forensic and clinical toxicology

Almost all of the above mentioned publications referred to bioanalytical methods for bioavailability, bioequivalence or pharmacokinetic studies. This field is of course very closely related to forensic and clinical toxicology, especially if only routine methods are considered. Therefore, it seems reasonable to base the discussion concerning method validation in toxicological analysis on the experiences and consensus described above and not to start the whole discussion anew. In the following, possible implications for forensic and clinical toxicology will be discussed.

Terminology

As already mentioned above, there are several sets of terminology in the literature. It is therefore strongly recommended to adopt, in principle, one of these sets for validation in forensic and clinical toxicology and add slight modifications, where it seems necessary. The definitions established by the ICH seem to be a reasonable choice as they are consensus definitions of an international conference and easily available on the homepage of ICH (www.ich.org).

Validation parameters

Selectivity (specificity)

During pharmacokinetic studies (therapeutic) drugs are usually ingested under controlled conditions. Therefore, there is no need to prove the ingestion of this drug. Due to this fact the selectivity evaluation can be based on the acceptability of precision and accuracy data at the LLOQ. This approach is quite problematic for forensic and clinical toxicology, where analysis is often mainly performed to prove ingestion of an (illicit) substance and, therefore, qualitative data are also important. Here, the approach to prove selectivity by absence of signals in blank samples makes much more sense. The confinement of Conference Report II [10] to only study one source of blank matrix for methods employing MS detection does not seem reasonable for toxicological applications because of the great importance of selectivity in this field. However, discussion is needed on how many sources of blank samples should be analysed and if this should depend on the detection method.

It seems also reasonable to also check for interferences from other xenobiotics that can be expected to be present in authentic samples (e.g. other drugs of abuse for methods to determine MDMA, other neuroleptics for methods to determine olanzapine). This can be accomplished by spiking these possible interferents at their highest expectable concentrations into blank matrix and checking for interferences after analysis. Another way to exclude interferences from such compounds is to check authentic samples containing these but not the analyte for interfering peaks. This latter approach is preferable, if the possibly interfering substance is known to be extensively metabolized, as it also allows exclusion of interferences from such metabolites, which are usually not available as pure substances.

Calibration model

The use of matrix-based calibration standards seems also important in toxicological analysis, in order to account for matrix effects during sample workup and measurement (e.g. by chromatographic methods). Consensus should be reached on how many concentration levels and how many replicates per level should be analysed. From our own experience six levels with six replicates each seems reasonable. Weighted calibration models will gen-

erally be the most appropriate in toxicological analysis, as concentration ranges of analytes in toxicological samples are usually much greater than in samples for pharmacokinetic studies. Homoscedasticity, a prerequisite for unweighted models, can however only be expected for small calibration ranges.

Accuracy (precision and bias)

There is no obvious reason to evaluate these parameters differently than has been described above. Due to the often higher concentration ranges, it might be reasonable to also validate the analysis of QC samples containing concentrations above the highest calibration standard after dilution or after reduction of sample volumes, as it has been described by Wieling et al. [6] and Dadgar et al. [5]. The latter has also described the use of QC samples with concentrations below those of the lowest calibration standard using greater sample volumes.

Limits

The same approaches and criteria as those described above under "Limits" could be used. All approaches have been described to a lesser or greater extent in international publications, especially for the determination of LOD. Nevertheless, it seems important to reach consensus on this matter at least for forensic and clinical toxicology, as reliable detection of a substance is one of the most important issues in toxicological analysis. At this point it must be stressed that for the estimation of LOD and LLOQ via a special calibration curve, the calibration samples must only contain the analyte at concentrations close to LOD and LLOQ. Use of the calibration curve over the whole range may lead to overestimation of these limits.

Stability

The biggest problems encountered during stability testing for bioanalytical methods in forensic and clinical toxicology is the fact that there is a great number of different sampling vessels. Furthermore, the anticoagulants used also differ. Both facts make it difficult to assess long-term stability, as the workload to analyse all possible combinations of vessels and anticoagulants is of course far to great. However, for some analytes relevant to forensic and clinical toxicology (e.g. cocaine) stability problems with different sampling vessels have been reported [19]. Therefore, the relevance of this parameter to forensic and clinical toxicology has to be discussed extensively. Agreement on a single type of vessel to use for sampling of toxicological samples would probably be the easiest solution. Another problem is the fact that storage conditions prior to arrival in the laboratory are not known. So this matter will also have to be addressed.

Recovery

Recovery does not seem to be a big issue for forensic and clinical toxicologists as long as precision, accuracy (bias), LLOQ and, especially, LOD are satisfactory. However, during method development one should of course try to optimize recovery.

Ruggedness

There is no obvious reason to treat this matter differently than described above under "Ruggedness (robustness)".

Measurement uncertainty

Measurement uncertainty, a parameter characterizing the dispersion of the values that could reasonably be attributed to the measurand (e.g. concentration), is considered an important concept in analytical chemistry [20]. It comprises many components and is generally reported as a standard deviation or as a confidence interval. However, measurement uncertainty was not explicitly addressed in any of the publications on bioanalytical method validation discussed in this review. One probable reason might be that measurement uncertainties of modern analytical methods are certainly small compared to the differences encountered between individual subjects in pharmacokinetic studies. Nevertheless, knowledge about the uncertainty of measurements can be very helpful or even important for the interpretation of bioanalytical data, especially in forensic toxicology. As bioanalytical methods are usually rather complex and assessment of the contribution of individual components on the combined uncertainty of the results would therefore be time consuming and costly, estimation of measurement uncertainty from validation data, especially precision and accuracy data, would be preferable. The more the design of validation experiments accounts for conditions during routine application of a method, i.e. the more individual components are comprised in one validation parameter, the better the estimation of measurement uncertainty from those parameters. Consequently, reproducibility data from interlaboratory experiments studies or intermediate precision data with M-factors ≥ 3 from intralaboratory experiments should adequately reflect measurement uncertainty of bioanalytical methods. However, further components not covered adequately by validation experiments might play a role for measurement uncertainty. A detailed guide on the quantification of analytical mea-

surement uncertainty has been published by EURACHEM/CITAC [20].

Conclusion

There are only a few principle differences concerning validation of bioanalytical methods in the fields of pharmacokinetic studies and forensic and clinical toxicology. Therefore, it seems reasonable to base the discussion on validation in the field of toxicology on the experiences and consensus already existing in the closely related field of pharmacokinetic studies for registration of pharmaceuticals and focus the discussion on those parameters, which are of special importance for toxicologists, i.e. selectivity, LOD, LLOQ and stability.

References

1. Karnes HT, Shiu G, Shah VP (1991) Pharm Res 8:421–426
2. Shah VP, Midha KK, Dighe S, McGilveray IJ, Skelly JP, Yacobi A, Layloff T, Viswanathan CT, Cook CE, McDowall RD, Pittman KA, Spector S (1992) Pharm Res 9:588–592
3. Hartmann C, Massart DL, McDowall RD (1994) J Pharm Biomed Anal 12:1337–1343
4. Dadgar D, Burnett PE (1995) J Pharm Biomed Anal 14:23–31
5. Dadgar D, Burnett PE, Choc MG, Gallicano K, Hooper JW (1995) J Pharm Biomed Anal 13:89–97
6. Wieling J, Hendriks G, Tamminga WJ, Hempenius J, Mensink CK, Oosterhuis B, Jonkman JH (1996) J Chromatogr A 730:381–394
7. Bressolle F, Bromet PM, Audran M (1996) J Chromatogr B 686:3–10
8. Causon R (1997) J Chromatogr B 689:175–180
9. Hartmann C, Smeyers-Verbeke J, Massart DL, McDowall RD (1998) J Pharm Biomed Anal 17:193–218
10. Shah VP, Midha KK, Findlay JW, Hill HM, Hulse JD, McGilveray IJ, McKay G, Miller KJ, Patnaik RN, Powell ML, Tonelli A, Viswanathan CT, Yacobi A (2000) Pharm Res 17:1551–1557
11. U.S. Department of Health and Human Services, Food and Drug Administration (2001) Guidance for industry, bioanalytical method validation http://www.fda.gov/cder/guidance/index.htm
12. Lindner W, Wainer IW (1998) J Chromatogr B 707:1–2
13. International Conference on Harmonization (ICH) (1994) Validation of analytical methods: Definitions and terminology ICH Q2 A
14. International Conference on Harmonization (ICH) (1996) Validation of analytical methods: Methodology ICH Q2 B
15. Vander-Heyden Y, Nijhuis A, Smeyers-Verbeke J, Vandeginste BG, Massart DL (2001) J Pharm Biomed Anal 24:723–753
16. Penninckx W, Hartmann C, Massart DL, Smeyers-Verbeke J (1996) J Anal At Spectrom 11:237–246
17. International Organization for Standardization (ISO) (1994) Accuracy (trueness and precision) of measurement methods and results: ISO/DIS 5725–1 to 5725–3. ISO, Geneva, Switzerland
18. NCCLS (1999) Evaluation of precision performance of clinical chemistry devices; Approved guideline. NCCLS, Wayne, Pa., USA
19. Toennes SW, Kauert GF (2001) J Anal Toxicol 25:339–343
20. EURACHEM/CITAC (2000) Quantifying uncertainty in analytical measurement http://www.eurachem.bam.de/guides/quam2.pdf

Accred Qual Assur (1997) 2:234–237
© Springer-Verlag 1997

Ilya Kuselman
Boris Anisimov
Avinoam Shenhar
Alex Lepek

Validation of a computer program for atomic absorption analysis

Received: 25 January 1997
Accepted: 14 March 1997

I. Kuselman (✉) · B. Anisimov
A. Shenhar
The National Physical laboratory of Israel, Danciger A. Bldg, Givat Ram, Jerusalem 91904, Israel
Tel: +972-2-6536534
Fax: +972-2-6520797
e-mail: freddy@vms.huji.ac.il

A. Lepek
Tech Projects, P. O. Box 9769, Jerusalem 91091, Israel

Abstract The approach to validation of a computer program for an analytical instrument as a component of the analytical method (using this instrument with the program) is discussed. This approach was used for validating a new program for atomic absorption analysis. The validation plan derived from this approach was based on minimising the influence of all steps of the analytical procedure on the analytical results obtained by the method. In this way significant changes in the results may be caused only by replacement of the previous program by the new one. The positive validation conclusion was based on the comparison of the results of the analysis of suitable reference materials obtained with the new program and with its precursor in the same conditions, and also on comparison of their deviations from the accepted reference values for these materials, with the corresponding uncertainties.

Key words Validation · Computer program · Analytical instrument · Measurement uncertainty · Atomic absorption analysis

Introduction

The validation of analytical methods is a well-known problem in the analytical community [1]. The international guidance for equipment qualification (EQ) of analytical instruments and their validation is in the development stage [2–4]. At this time validation of computer systems for analytical instruments is less elaborated [5–7]. The term "computer system" comprises computer hardware, peripherals, and software that includes application programs and operating environments (MS-DOS, MS-Windows and others) [5, 6]. Since programs, software and the whole computer system are elements of the instrument used by the analyst according to the analytical method, successful validation of the method as a "black box" [8] means successful validation of the instrument, computer system, software and programs. On the other hand, the same instrument may also be calibrated and validated as a smaller (included) "black box" [9], with a corresponding validation conclusion about the computer system and its elements. Next, the computer system is again a smaller included "black box" [10, 11] etc. It is like a matreshka – a Russian wooden doll with successively smaller ones fitted into it. Therefore the approach to the validation of a computer program as a validation of a component of the analytical method is sound. In the framework of this approach, not all the method validation parameters (performance characteristics [10, 11]) may be relevant for the program validation. For example, the validation parameter "accuracy" (bias) is meaningful in this context, while the "selectivity", "specificity" or "ruggedness" are not. The bias may be evaluated as the difference between the results of the analysis of suitable reference materials obtained with the new program and with its precursor in the same conditions, and also by evaluation of their deviation from the accepted reference values for these materials in comparison to the corresponding uncertainties [12]. Certainly, the pro-