EPMA Journal (2010) 1:619–626
DOI 10.1007/s13167-010-0053-y

REVIEW ARTICLE

The role of laboratory medicine in healthcare: quality requirements of immunoassays, standardisation and data management in prospective medicine

Thomas Waerner · Dietmar Thurnher · Kurt Krapfenbauer

Received: 14 September 2010 / Accepted: 21 October 2010 / Published online: 2 December 2010

Abstract In the last 10 years, the area of ELISA and protein-chip technology has developed and enthusiastically applied to an enormous variety of biological questions. However, the degree of stringency required in data analysis appears to have been underestimated. As a result, there are numerous published findings that are of questionable quality, requiring further confirmation and/or validation. In the course of feasibility and validation studies a number of key issues in research, development and clinical trial studies must be outlined, including those associated with laboratory design, analytical validation strategies, analytical completeness and data managements. The scope of the following review should provide assistance for defining key parameters in assay evaluation and validation in research and clinical trial projects in prospective medicine.

Keywords Quality requirements · ELISA · Multiplex assay · Validation · Laboratory workflow · Quality guidelines

D. Thurnher is National Representative of EPMA in Austria.

T. Waerner
Quality & Compliance, Cell & Molecular Biology,
Boehringer-Ingelheim RCV GmbH & CoKG,
Dr. Boehringer-Gasse 5-11,
A-1121 Vienna, Austria
e-mail: Thomas.waerner@boehringer-ingelheim.com

D. Thurnher
Department of Cranio-Maxillofacial and Oral Surgery,
Medical University of Vienna,
Waehringer Guertel 18-12,
A-1090 Vienna, Austria
e-mail: Dietmar.thurnher@meduniwien.ac.at

K. Krapfenbauer (✉)
Process Science, In-Process-Control,
Boehringer-Ingelheim RCV GmbH & CoKG,
Dr. Boehringer-Gasse 5-11,
A-1121 Vienna, Austria
e-mail: Kurt.krapfenbauer@boehringer-ingelheim.com

Introduction

ELISA multiplex protein-chip technology and imaging techniques, are powerful means of rating a lot of analytical data in the frame of a single experiment. This analytical information is used to understand the nuance of the protein expression profile of a biological system, and in many cases it is the basis of comparison of two or more sample sets. Yet the technical difficulty and high cost of data production, associated with highly time-consuming data analysis, has contributed to a position where a poor laboratory workflow is common. Many experiments have a low number of analytical and/or biological replicates, and user often assume that multiple estimates rated by a single experiment provide a substitute for experimental replicas. The reproducibility of techniques used, as assayed by regression analysis, coefficient of variation or other variance estimation techniques [1] is typically not reported. Power analysis, which can be used to infer the number of samples that should be analysed to discover a statistically significant result [1–3], are rarely undertaken. Weak design of experiments, particularly in a field where technical challenges remain in the production of high quality data, can make it difficult or impossible to determine if differences reported between two or more sample sets are likely to reflect variation in a biological system or are solely analytically derived.

The complexity of an analytical validation should reflect the aim of the analysis and thus has to be in accordance with the intended use. In other words, a validation of a method for a research project has to be scientifically defendable whereas a method designed for tracking critical parameter during development of a

manufacturing process or ensuring product quality parameters within specified limits has to be validated in accordance with cGMP standard [4]. Therefore, validation can be seen as the process of demonstrating that an assay is suitable for its intended use [4].

The scope of the following proposals and examples should provide assistance for defining good practice for assay performance in research and early phase clinical trial projects. The process by which a specific bioanalytical method is developed, evaluated and finally validated for it's intended use can be divided into (1) reference standard preparation, (2) bioanalytical method development and establishment of assay procedure, and (3) application of the validated bioanalytical method for routine analysis and release for the analytical run and/or batch. Commercially available tests have not necessarily undergone all of these steps. Asking the vendor for additional data on assay development and evaluation may significantly and easily improve the analysts knowledge about assay properties.

For analysis of single compounds in complex matrices like different human fluids, specificity (selectivity) is a prerequisite for valid results. According to the ICH Guideline Q2R, specificity is the ability to assess unequivocally the analyte in the presence of components which may be expected to be present. Typically these might include impurities, degradants, matrix, etc.

Furthermore, using test samples with known concentration or generating samples with spiked standard substance allows to describe the difference between the value obtained by the test method and the theoretical value which is defined as the assay accuracy.

For ELISA-, and protein-chip-assays, but also for other bioassays, it is strongly recommended that reference material is identical to the sample material and reference matrix and sample matrix have identical compounds or at least behave similar. If possible, an adequate number of control samples should be prepared and stably stored early during assay development. If tested with each assay run, this control provides great advantage by monitoring the assay performance. It is a valuable tool to track assay performance parameters like the assay variability over time. The random assay variability gives a first, rough estimate on the QL (Quantitation Limit) which is the lowest level of analyte that can be measured with sufficient precision and accuracy. For this purpose 6 times the standard deviation is a rough estimate for the QL which should be analytically proved. All values below the QL should be reported as "below QL", or below the concentration value corresponding to QL. No scientific conclusion should be taken from comparing of results that are below the QL. Assay variability should be investigated and reported as the assay precision. The precision of an analytical method expresses the closeness of agreement between a series of measurements obtained from multiple sampling of same homogeneous sample and may be reported as the coefficient of variation CV%. Taking into consideration that most studies are performed on more than 1 day, the effects of random events to the analytical procedure may be tested to ensure comparable test results between e.g. different test starts, different analysts, different lots of antibodies or different equipment. The interval between the upper and lower concentration of analyte in the sample where the procedure has a suitable level of precision and accuracy should be specified as the range of the method.

Control samples are also important to define limits for assay acceptance criteria. If the actual assay differs more than 3 times the standard deviation from the historical mean, it may be investigated if an analytical error has occurred and if an error is identified, the assay may be repeated without considering the initial results.

Monitoring of control samples make it possible to identify a bias (difference) in the results over time or a bias that appears suddenly between two assays. The latter might be caused by the change of a reagent, equipment or analyst without performing material bridging study or sufficient training of the new analyst. There are many arguments why such a bias should be avoided within a study. But if it happens the back calculation of sample results with the tracked bias of the control samples may save the study. This is the case if the analyte is instable and samples can not be stored and therefore the testing can not be repeated.

Investigation of change of sensitivity across a plate is important as usually multiple wells of a multiwell-plate are used for the testing procedure. Therefore the uniformity of the assay performance over the plate should be investigated by a single dose that is used across the entire plate. Therefore, a sample concentration within the reference standard curve should be chosen that represents the greatest sensitivity to change in dose. The pipetting scheme and the output detection of each individual well should be processed in the same way as planned for the test. The results may be shown in a 3D Uniformity trial plot [5], see Fig. 1.

If deviations from uniformity are detected, the plate layout can be designed to minimise plate effects. In the shown example, the right side of the plate shows reduced sensitivity compared to the left side. In such a cases the cause should be investigated. One possible cause may be the time limitation of a luminescent readout. In such a case the shortening of the reading time for each well increases the assay variability but reduces the bias across the plate. Furthermore an appropriate plate design that spreads the duplicates across the plate may be also useful to minimise

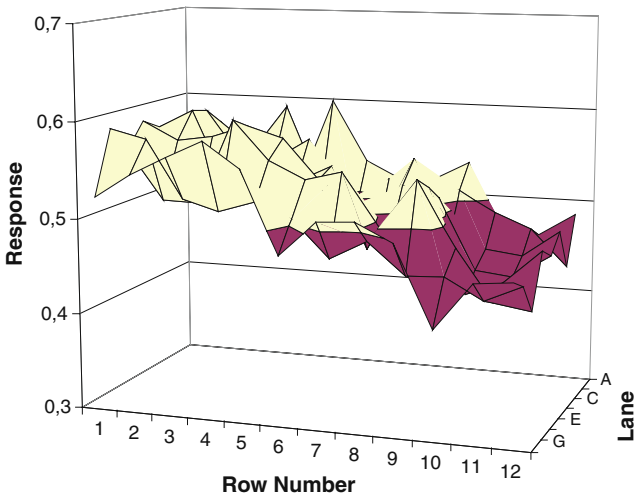

Fig. 1 Example of a 3D uniformity plot of sensitivity across a plate. Each well of a 96 well plate contains the same amount of analyte. Please note the bias of the readout response from left to right. This effect was caused by an instable readout signal and slow well scanning row by row, starting at row 1 lane A, ending with row 12 lane F

the bias of samples located to different sites of the plate [5]. If a bias occurs, a suitable template design minimises the bias, mainly by transforming the bias into assay variability. Assay variability can be handled statistically e.g. by an increased number of replicates that are spread over the plate.

Assay variability, biological variation of the samples and small differences in sample pre-treatment may lead to variability of the assay results. To consider this fact, it is suggested not only to report the mean of the individual replicates as result but also report the consensus interval (e.g. CI 95%) or the standard deviation for improved interpretation of the data. If the assay results are not normally distributed, the geometric mean of the replicates, rather than the arithmetic mean should be reported [6]. Assay performance parameters are given in Table 1.

Differential display and biomarker discovery

ELISA, multiplex assay and protein-chip techniques are widely used for the determination of differentially expressed proteins, including biomarkers. Multi screening techniques can be used in a hypothesis-independent manner (change from hypothesis to discovery driven research), making them attractive for this purpose. Whilst statistical tests are increasingly applied to expression data, proteins are frequently published as differentially expressed on the basis of a two-fold or greater expression difference. Such conclusions ignore the analytical and biological variation inherent to any laboratory and the samples under study. It is also not infrequent to see proteins described as differentially expressed from the use of univariate statistical tests (e.g. Student's *t*-test), but where the normal distribution of the data has been assumed but not tested. This is of great concern as expression data are typically not normally distributed and requires transformation before many statistical tests can be applied [1, 3]. After appropriate statistical analysis, it may come to pass that a two-fold expression difference is shown to be significant for a particular protein. However, it is only the detailed analysis of expression data, involving data normalisation, appropriate transformation, determination of the inherent variance and the use of suitable uni- and multivariate statistical test, that this can be resolved.

Analytical incompleteness of multiplex protein expression/determination analysis

Analytical incompleteness refers to a phenomenon where a technique used for the analysis of a complex mixture of peptides/proteins may only yield information for a fraction of relevant peptides/proteins in any single analytical run. For example, it has been observed that two replicate analysis of a multiplex protein-chip experiment will produce two sets of different expression profiling with ~65% overlap [7, 8]. Thirty-fife percent of the protein in the second analysis are likely to be novel compared to the first. A third replicate analysis is likely to yield a set of identification that has 80% overlap with those from the first two analyses, but with 20% new identification. Because of the differences in proteins seen *per* run, it has been estimated that 10–12 analysis may be necessary before a near complete of protein identities is rated from a single complex sample [7, 8]. This phenomenon has a substantial impact on the use of protein arrays for qualitative biomarker discovery experiments, as the presence or absence of a protein in a particular run may reflect analytical incompleteness instead of true differences between samples. Accordingly, the comprehensive comparison of two or more proteome analysis by these approaches will require great care and high numbers of replicates [9]. It should finally be noted that analytical completeness is also inherent to the technique, where it can arise from inconsistent sample extraction and amplification, leading to difference in the protein expression determined in western blot (WB), ELISA etc. [10]. However the paradigm of proteomics via affymetrix typically identifies proteins or genes of interest only after statistical expression analysis, making this a less pressing issue.

In the content of the above concerns, we recommend initiate a process to develop a set of minimum guidelines for the field of ELISA, and multi-protein-chip assay, which could assist in the execution of protein determination experiments and in the improving of the quality in the data.

Figure 2 demonstrates evaluation, validation and test validation processes for immunoassays.

Table 1 Typical assay performance parameters addressed in assay validation to make the assay scientifically defendable

Assay performance parameter	Outcome / benchmark	Examples
Specificity	Detection of the analyte. No detection of matrix components.	No cross reactivity with other components that are expected in the sample matrix.
Linearity	If applicable, linear range of the standard curve R^2>0,98	5 dilution points of the standard curve show a linear dose response relation.
Range	Sample concentration where linearity, accuracy and precision are within the expected limits. The lowest valid concentration of the range is the Quantitation Limit	Results from accuracy and precision can be used.
Accuracy	80–120% of spiked value	Sample matrix spiked with 5 known concentrations of standard.
Intermediate precision	RSD% <25%	2 different analysts, 2 different days, 2 different readout systems or batches of antibody
Robustness	Investigation of assay sensitivity to possibly occurring influences.	Incubation time and temperature, performance of plate washing, equipment

The shown benchmarks represent an estimate for multiple types of immunological assays and strongly depends on the assay method and the intended use

Recommended workflow: assay validation protocol for immunoassays (ELISA and multiplex protein-chip assay) proposed for prospective medicine

Prevalidation / evaluation

Order 2 kits from 2 different suppliers for the analyte-specific assay which has to be validated. If necessary, get additional aliquots of the lot-specific standard. If a WHO standard preparation of the analyte studied is available, order as well. Alternatively, get a large stock of a stable preparation (lyophilised) of analyte from independent suppliers.

a) Prepare a standard curve with the kit-standard spiked in one human plasma and in horse plasma; run the spiked plasma samples along with kit standards in duplicate on one plate.
b) Run 25 different human plasma samples from different donor in triplicates.

If the majority of data (>80%) fulfill a % CV criteria of below 20% (data from a & b) and a constant spike recovery (data from a—after accounting for endogenous level in human plasma) above 50% over the concentration range tested continue with the validation.

Validation

1. Depletion of endogenous analyte-level => assessment of specificity of baseline levels
   - If the analyte levels are in agreement with the literature data skip immunodepletion and continue with point 2.
   - Take the capture antibody from the assay you want to validate.
   - Add the antibody to the 10 human plasma samples from different donor at a concentration of 20 μg/ml.
   - Incubate the plasma samples with the antibody overnight at +4°C on ice.
   - Run the undepleted and depleted plasma samples in triplicate.
   - If the baseline levels can be depleted by two-third or more continue with point 2.
   - If the baseline levels can not be depleted by two-third or more discard the assay.
2. Different human plasma => assessment of major matrix effects
   - Determine the analyte-level of 10 different human plasma samples from different donor (the results from the prevalidation can be taken for selecting the plasma in order to cover a maximum range).
   - Spike the standard-protein in two different concentrations to the individual plasma (3× and 10× of the mean endogenous level).
   - Run this unspiked and spiked plasma samples in triplicate.
   - If the %CV are below 20% and the spike recovery constant over the individual patient samples tested and exceeds 50%, continue with point 3.
3. Variability and LLoQ
   - Prepare a standard curve with the kit-standard spiked in one human plasma and in horse plasma identical to point 1. The human plasma with the

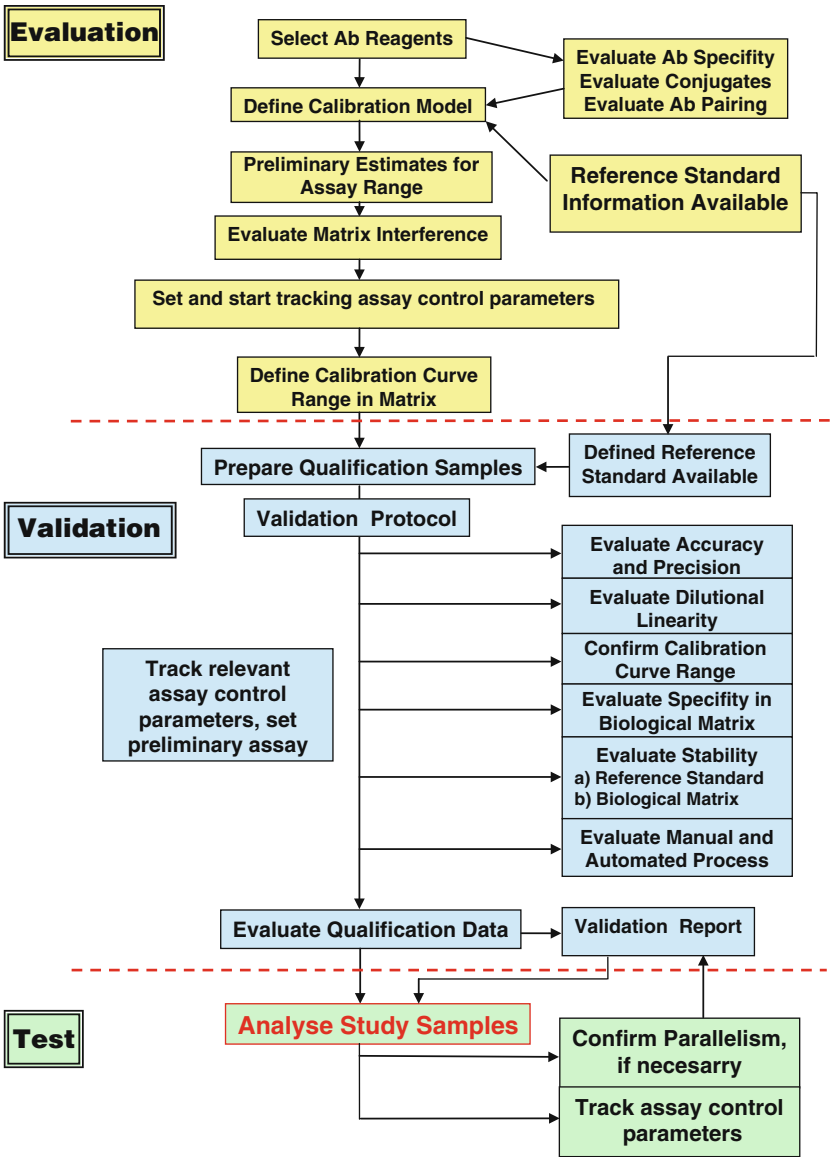

Fig. 2 A scheme of evaluation: Typical assay performance parameters addressed in assay validation and test validation processes for immunoassay, taken modified from [11]. The shown benchmarks represent an estimate for multiple types of immunological assay and strongly depends on the assay method and the intended use

lowest endogenous signal identified in the prevalidation should be used.

- Run this standard curve in duplicate/plate on two different plates; this should be done independently by two different operators on two different days (the results of the prevalidation can be also included in the validation).
- From this results assess precision and accuracy and derive LLoQ (also mentioned as QL = Quantitation Limit) is defined by the spike concentration where the CV is below 20% and the recovery is still in a constant range.

4. In-run controls

- From the 25 different human plasma samples (prevalidation), define three with different concentrations (>3 fold difference between the concentrations) as assay controls; if the plasma samples do not display analyte levels that span an appropriate range the analyte has to be spiked to one plasma in order to achieve three different concentrations.
- Include four replicates of these three control plasma samples per plate (from part 3 above) at different positions on the plate to detect the mean and the range of the controls out of 16 replicates.
- These controls have to be included in duplicate on all assay plates where clinical samples are run (so-called *in-run controls*); 5/6 of the in run-controls need to be within 20% of the mean of the 16 results.

5. Assessment of independent standard
   - An independent standard preparation of the analyte (if available, WHO standard) should be spiked at 5 different concentrations into horse plasma and assessed in duplicate over the four plates used in 3. Average and % CV should be derived. Any new lot should be tested with identical spikes and should only be released for clinical sample analysis if independent standard spikes are in agreement with the one determined during validation.

Recommendation for data management

Laboratory workflow and data analysis for ELISA analysis and multiplex-based experiments:

- The laboratory workflow must be provided and must include details of the number of biological and analytical replicates. Only one biological/analytical replicate is not acceptable. For clinical samples, it is highly desirable that a power analysis predicting the appropriate sample size for subsequent statistical analysis of the data is carried out.
- For protein expression studies, summary statistics (mean, standard deviation) must be provided and results of statistical analysis must be shown. Reporting fold differences alone is not acceptable. The report must include the following: method of date normalisation, transformation, missing value handling, the statistical tests used, the degrees of freedom and the statistical package or programme used. Where biologically important differences in protein expression are reported, confirmatory data (e.g. Western Blot (WB), ELISA) are desirable.
- For biomarker discovery/validation studies, the sensitivity and specificity, the positive predictive value (PPV), negative predictive value (NPV), likelihood ratio (LR) and odds ratio (OR) of the biomarker(s) should be provided wherever possible. It is desirable that receiver operator characteristic (ROC) curves and areas under the curves are given.

Protein identification and characterisation

- The method(s) used to measure the protein expression data must be described.
- The name and version of the programme used for database searching, the values of critical search parameters must be provided.
- For experiments with large protein expression data sets, estimates of false positive rates are required (e.g. through searching randomised or reversed sequence database). This information should be provided as supplementary material.
- When post translation modification (PTM) of proteins are reported, the amino acids sequence that matches the unique peptide sequence of a particular isoform must be provided.

Bioinformatics

- Where a report describes an academic database or software, it must be either freely accessible *via* the internet, intranet or downloadable and the access options must be provided. This also applies to commercial software or databases.

Normalisation

- *Default normalisation*: Values below 0.01 were set to 0.01. Each measurement must divided by the 50.0th percentile of all measurements in that sample. Each protein analyte must divide by the median of its measurements in all samples. If the median of the raw values was below 10 then each measurement for that protein was divided by 10. This normalisation procedure must used to calculate the standard correlation of all samples.
- *Normalisation to control samples*: Values below 0.01 were set to 0.01. Each measurement was divided by the 50.0th percentile of all measurements in that sample. Treated samples must normalise against the medium of samples of the same time point. Each measurement for each biomarker protein in those specific samples should divide by the median of that protein's measurements in the corresponding control samples. For all calculations within or in between samples this normalisation procedure should chosen.

Statistics

- Due to difficulties either in sample preparation, in protein preparation or in assay or protein-chip hybridisation the amount of replicas varied from zero to six. Thus implicating different optimal statistical tests were necessary for the various settings.

Filter by Control Strength

- Control strength (CS) is a synthetic control value that resulted from the normalisation steps. In

common case, it is equal to the median of the per assay normalised expression values of the control samples. Measurements with higher control strength are relatively more precise than measurements with lower control strength. If all control strength values are plotted against the standard deviation of the normalised value, the best cut off to filter data is where the curve flattens out (where the measurement for the data becomes more reliable).

Log transformation

- Log transformation can minimise the impact of outliers with high signals (essentially they make non-normal distributions look more normal-like). Data should be interpreted in log of ratio mode because parametric tests assume that means of the population under study are normally distributed (Gaussian distribution). All statistical tests must apply to the distribution of natural logs of the ratios for each protein.

One sample *t*-test

- The one sample *t*-test, determines the likelihood if the average ratio in the log replicates interpretation is significantly different from 1.0. A filter (*P*ÿ.05) will be applied to the *P* values to determine the statistical significance of each protein's differential expression.

1-way ANOVA

- One-way analysis of variance (ANOVA) tests allow determining if, in our case treatment of patient with drugs, has a significant effect on protein expression behavior across the groups untreated and treated at a specific time point. A Welch *t*-test (variances assumed not to be equal) using GEM variances with p ÿ 0.05 should be performed. It will be not possible to apply this approach to the comparison of all samples, due to its stringency.

MTC: Multiple Testing Corrections

- To avoid type 1 error (false positives) which occurs when the biomarker/analyte is not differentially expressed and the analysis concludes that this is significant, Benjamini and Hochberg False Discovery Rate should used. The purpose of a multiple testing correction is to keep the overall error rate/false positives to less than the user specified *p*-value cutoff, even if several biomarkers/analytes are being analysed. In datasets, where samples from different cohorts were compared, no MTC could be applied because it will be too restrictive.

Sample Specificity

- Before starting the analysis, the samples should checked for quality by measuring their similarities within a given samples cohort. The "Find Similar Samples" allows running a comparison between a target sample and a specified group of samples. The algorithm uses a Spearman correlation, which takes the relative ranks of the raw expression values (as opposed to the normalised values).

Protein Ontology Classification

- Ontology comprises a set of well-defined terms with well-defined relationships. The structure itself reflects the current representation of (biological) knowledge as well as serving as a guide for organising new data.
- Out of the three described classifications (molecular function, process and subcellular location) we will chose molecular function to classify our proteins. Molecular function is defined as the biochemical activity (including specific binding to ligands or structures) of a protein product. This definition also applies to the capability that a protein product (or protein product complex) carries as a potential [12].

Recommendations and outlook

Immunoassays are applied in such important areas as the quantitation of biomarker molecules in prospective medicine which indicate disease progression or regression, and antibodies elicited in response to treatment with therapeutic drug candidates. Currently available guidance documents dealing with the validation of bioanalytical methods address immunoassays in only a limited way. In the course of our review we present recommendations for specific aspects of immunoassay characterisation and validation. Immunoassay calibration curves are inherently nonlinear, and require nonlinear curve fitting algorithms for best description of laboratory data. Demonstration of specificity of the immunoassay for the analyte or biomarker of interest is critical because most immunoassays are not preceded by extraction of the analyte from the matrix of interest. Since the core of the assay is an antigen-antibody reaction, immunoassays may be less precise and less specific leading sometime to false positive results. Criteria for accuracy (mean bias) and precision, both in pre-study validation experiments, and in

the analysis of in-study quality control samples, should be more lenient than for other assay. Our recommendations for immunoassay validation are presented in the hope that their consideration may result in the production of consistently higher quality data from the application of these methods in predictive, preventive and personalised medicine.

References

1. Hunt KJ, Lehman DM, Arya R, Fowler S, Leach RJ, Göring HH, et al. Genome-wide linkage analyses of type 2 diabetes in Mexican Americans: the San Antonio Family Diabetes/Gallbladder Study 2005. Diabetes. 2005;54(9):2655–62.
2. Molloy MP, Brzezinski EE, Hang J, McDowell MT, VanBogelen RA. Overcoming technical variation and biological variation in quantitative proteomics. Proteomics. 2003;3(10):1912–9.
3. Karp JM, Friis EA, Dee KC, Winet H. Opinions and trends in biomaterials education: report of a 2003 Society for Biomaterials survey. J Biomed Mater Res A. 2004;70(1):1–9.
4. ICH Guidelines: Q2(R2), Validation of analytical procedures: text and methodology. 2005; Q6B, Specifications: test procedures and acceptance criteria for biotechnological/biological products. 1999.
5. United States Pharmacopeia <1032> Design and development of biological assays; draft version 2010, public in 2011.
6. United States Pharmacopeia <1034> Analysis of biological assays; draft version 2010, public in 2011.
7. Durr PA, Eastland S. Use of web-enabled databases for complex animal health investigations. Rev Sci Tech. 2004;23 (3):873–84.
8. Lui H, Qiu T. Knowledge discovery in database and its application in clinical diagnosis. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi. 2004;21(4):677–80.
9. Orchard S, Hermjakob H, Taylor CF, Potthast F, Jones P, Zhu W, et al. Further steps in standardisation. Report of the second annual Proteomics Standards Initiative Spring Workshop (Siena, Italy 17–20th April 2005). Proteomics. 2005;5(14):3552–5.
10. Challapalli KK, Zabel C, Schuchhardt J, Kaindl AM, Klose J, Herzel H. High reproducibility of large-gel two-dimensional electrophoresis. Electrophoresis. 2004;25(17):3040–7.
11. Findlay JW, Smith WC, Lee JW, Nordblom GD, Das I, DeSilva BS, et al. Validation of immunoassays for bioanalysis: a pharmaceutical industry perspective. J Pharm Biomed Anal. 2000;21(6):1249–73.
12. Ashburner J, Friston KJ. Voxel-based morphometry—the methods. Neuroimage. 2000;11(6 Pt 1):805–21.