Journal of Pharmaceutical and Biomedical Analysis 55 (2011) 869–877

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

Validation of immunoassay for protein biomarkers: Bioanalytical study plan implementation to support pre-clinical and clinical studies

Marie-Anne Valentin[a,*], Shenglin Ma[b], An Zhao[b], François Legay[a], Alexandre Avrameas[a]

[a] Department of Marker Localization and Assays, Novartis Institute for Biomedical Research, CH-4056 Basel, Switzerland
[b] Department of Marker Localization and Assays, Novartis Institute for Biomedical Research, Shanghai 201203, China

ARTICLE INFO

Article history:
Received 27 October 2010
Received in revised form 15 March 2011
Accepted 22 March 2011
Available online 29 March 2011

Keywords:
Immunoassay
Endogenous biomarker
Bioanalytical study plan
Feasibility study
Validation
Parallelism

ABSTRACT

Biomarkers have emerged as an important tool to optimize the benefit/risk ratio of therapeutics. The scientific impact of biomarker studies is directly related to the quality of the underlying data. It is therefore important that guidance be established for validation of assays used to support drug development. This paper specifically focuses on validation of immunoassay for protein biomarker to support pre-clinical and clinical studies. Therapeutics (small- and macro-molecules) and their respective target/ligand are out of scope. This paper describes the implementation of a bioanalytical study plan for the validation of immunoassays to support decision-making biomarkers and biomarker selection during preclinical and clinical studies. It establishes the complete operating procedure as well as the parameters and their respective acceptance criteria and defines milestones and decision points to be followed during the assay validation that should result in high quality bioanalytical data in a limited timeframe and with reduced costs. The bioanalytical study plan can be applied to the validation of a wild range of immunoassay technology such as monoplex ELISA, automated analyzer, multiplex assays or cutting edge technology.

Before any validation, a feasibility study is performed to assess the performance of the immunoassay using biological samples which should mimic the clinical population. The feasibility study addresses the likelihood that an assay will be able to achieve its intended purpose with parallelism being the most critical element (milestone 1). At the end of the feasibility study, a decision is taken to either continue with the validation or change the assay (milestone 2). The milestone 3 consists of the establishment of the nominal value of quality control to be used during the validation. The quality controls used to validate an assay should preferentially be prepared using neat (non-spiked) biological matrix (ideally derived from the specific trial population). The last milestone (milestone 4), the formal validation, includes demonstration of the assay performance meeting accuracy and precision acceptance criteria within (intra-run) and between (inter-run) validation runs for each QC sample. Validation also includes the assessment of stability of the protein biomarker in the biological matrix.

It is recognized that the extent of the validation should be correlated to the intended use of the data and the assay acceptance criteria should take into consideration the study objective(s), nature of the methodology and the biological variability of the biomarker.

1. Introduction

The challenge in drug development is to develop efficacious treatments with limited safety issues; to this end, biomarkers are being used to optimize the benefit/risk ratio of therapeutics. Biomarkers can be extremely valuable in drug development to assess the mechanistic effects of a drug (i.e. does a drug hits its target?), to select or monitor dose, to monitor the effects of a drug on a disease or on safety parameters, or to select patients for treatment.

The utility of biomarkers to support decision making in clinical studies, or to conduct exploratory studies, is directly related to the quality of the underlying bioanalytical data. It is therefore important that the principles for the validation of assays be established. There are no specific regulations on bioanalytical method validations but the Food and Drug Administration (FDA), other agencies and industry task forces have developed guidelines for method validation. Bioanalytical workshops were conducted over the past two decades. The first bioanalytical validation workshop was organized in 1990 by the American Association of Pharma-

Abbreviations: AAPS, American association of pharmaceutical scientists; ACC, accuracy; STD, calibration standard; LLOQ, lower limit of quantification; MRD, minimum required dilution; CV, precision; QC, quality control; ULOQ, upper limit of quantification.

* Corresponding author at: Novartis Pharma AG, Fabrikstrasse 10-4.40.07, CH-4056 Basel, Switzerland. Tel.: +41 61 696 79 79; fax: +41 61 696 74 87.
E-mail address: marie-anne.valentin@novartis.com (M.-A. Valentin).

0731-7085/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2011.03.033

ceutical Scientists (AAPS), USFDA, International Pharmaceutical Federation, Health Protection Branch and Association of Analytical Chemists. The workshop defined the parameters essential to ensure the acceptability of the performance of the bioanalytical method: accuracy, precision, selectivity, sensitivity, reproducibility and stability [1,2]. A second workshop in 2000 reemphasized the importance of standard curve and quality controls acceptance criteria. It also reemphasized that it is not necessary to have 100% recovery but it is important to have reproducible and consistent recovery [3]. The purpose of such document is not to provide direction on how to perform a validation but rather to present a discussion of the characteristics for consideration during the method validation. The guidelines focus on bioanalytical method validation to support pharmacokinetic studies contributing to the confusion regarding biomarker data [4–8]. As opposed to pharmacokinetic assays, biomarker methods (and in particular those related to immunoassays) commonly experience the following issues: the absence of suitable reference material, the employment of unique analytical reagents and analytical platforms, the presence of endogenous biomarker in biological matrix, and a plethora of disease-specific effects. Because of this, direct application of regulatory guidelines for either pharmacokinetics assays [9–11] or diagnostic development [12] is often precluded. The AAPS and Clinical Ligand Assay Society cosponsored a Biomarker Method Validation Workshop in October 2003 purpose [13]. It was concluded that biomarker assays must be validated in a manner that provides reliable and accurate data as basis for decision-making. The results of discussion were summarized in a position paper published in 2006 which recommends an iterative, fit-for-purpose approach for biomarker assay validation. The key elements for validation were defined as sample stability, preparation of calibration standard and quality controls, setting target and dynamic range with appropriate calibration curve-fitting, selectivity, precision and accuracy [14,15]. Although the aim of the paper was to stimulate more discussion and to foster consensus in the field of biomarker, regulatory guidance are still missing and few publications for validation of biomarkers were published since 2006 [16–19]. This is the reason why we undertook the effort to develop specific operating procedures for bioanalytical method validation to support biomarker studies.

This article establishes a bioanalytical study plan for validation of bioanalytical methods for protein biomarkers with emphasis on immunoassay techniques. Biotherapeutics (small- and macro-molecule drugs) and their respective target, measured by ligand-binding assay, mass-spectrometry or chromatographic-based systems are out of scope as the AAPS and FDA already published consensus reports on their validation [20–22]. There is also assumption that small-molecule biomarkers (such as steroids) measured by LC–MS will likely fall into the definitive quantitative assay category and will meet the guidelines for pharmacokinetic assays [23]. As the calibration curve is prepared in a surrogate matrix with a well-characterized reference material, there is no reason to relax the accepted pharmacokinetic acceptance criteria [23]. On the contrary, immunoassays used to measure biomarkers are relative quantitative assays as the reference material used to prepare the calibration curve is not truly representative of the endogenous biomarker and thus justify the use of “relaxed” acceptance criteria [14,16,17].

We view biomarker assay validation as a continuous process that allows the quantitative analysis of a specific biomarker, by identifying and monitoring sources of analytical error. We define key elements and their respective accuracy and precision acceptance criteria and propose milestones and decision points to be followed during the assay validation that should result in high quality bioanalytical data in a limited timeframe and with reduced costs. Amongst the key elements to be assessed during the validation [14], we put the emphasis on parallelism. Most of the biomarkers are measured using relative quantitative assays as the reference material used to prepare the calibration curve is not truly representative of the endogenous biomarker [14,16,17]. As a result, assumptions are made regarding assay accuracy. Such assumptions are tested in the parallelism experiment which should demonstrate that the concentration–response relationship of the biomarker in the matrix is sufficiently similar to that in the assay buffer. If there is no proportionality between the endogenous form and the reference standard, the method cannot be defined as a relative quantitative assay but rather as a quasiquantitative. Examples of such assays include quantitative RT-PCR and immunogenicity [14,16,17]. Parallelism assessment is very often missing from the publication for immunoassay validation or mistaken with selectivity [24–28]. Therefore, parallelism is a key element and should be assessed before validation is initiated. Prior to validation, we suggest performing a feasibility study to assess the performance of the kit or the developed immunoassay using biological matrix which should mimic the clinical population. During the feasibility study, parallelism, selectivity and reproducibility are assessed. The feasibility study addresses the likelihood that an assay will be able to achieve its intended purpose. Validation should only be performed if feasibility study meets the defined criteria.

The next milestone is the preparation of the quality controls (QCs). QCs are distinct from calibration standards (STDs) and are used to establish the working range. QCs used to validate an assay should be prepared using the biological matrix (ideally derived from the specific trial population). QCs prepared in buffer are not recommended as they do not represent the endogenous biomarker in the authentic samples. Accuracy and precision data of QCs prepared in the biological matrix are believed to represent more realistically the assay performance of clinical samples. As opposed to position papers previously published [14,15,18], we give the priority to neat (non-spiked) rather than spiked sample as they truly reflect the trial population. This is especially true for the QC targeting the lower limit of quantification where the assay variability is increased [18,19]. The true/nominal value of the QCs is determined from multiple runs performed without defined acceptance criteria, used to assess the assay variability.

The last milestones, the formal validation, include demonstration of the assay performance meeting accuracy and precision acceptance criteria within (intra-run) and between (inter-run) validation runs for each QC sample. Assay accuracy and precision are performance characteristics that describe the magnitude of errors associated with repeated measurements. QC samples are used to assess the ability of the assay to measure the biomarker of interest for its intended use, allowing one to distinguish assay variability from inherent differences amongst samples. Validation also includes the assessment of the stability of the protein biomarker in the biological matrix.

While human biomarker studies fall outside the scope of GLP, we recommend performing biomarker assay validation in compliance with GLP, meaning that the standard process for GLP validation should serve as a framework [3,8,11]. As suggested by Lee et al. [14], bioanalytical method validation should be tailored to meet the intended purpose of the study. We have translated the adoption of such a practical and “fit-for-purpose” approach into two levels of validation corresponding to two types of biomarker studies: decision-making biomarkers and biomarker selection. The first level of validation applies to biomarkers considered primary or secondary endpoints of a clinical study protocol and used to support internal decision making in clinical studies. The second level of validation applies to biomarkers considered exploratory endpoints and used to support exploratory studies or objectives. The rigor of biomarker method validation increases as the biomarker data are used for increasingly critical decision-making.

The recommended bioanalytical study plan can be applied to several immunoassay platforms including but not limiting to classical monoplex sandwich ELISA, automated analyzer (such as Elecsys or Immulite platform), multiplex planar array assays (such as MesoScale Discovery or SearchLight platform), multiplex suspension microbead assays (such as Luminex platform) and cutting edge of the technology (such as the flow-based single-molecule counting Erenna® platform from Singulex).

This paper provides a bioanalytical study plan for the validation of immunoassays to support decision-making biomarker and biomarker selection during preclinical and clinical studies. It establishes the appropriate procedure as well as the parameters and their respective acceptance criteria for the validation of an immunoassay to support biomarker studies. By defining milestones and decision points, we are aiming to optimize the production of high quality bioanalytical data.

Throughout the document, critical element and specific recommendations are highlighted in bold.

2. Method development

During the development of a biomarker assay, the following parameters should be defined: assay format, antibody(ies), plate type, detection system, critical assay reagent and stability, standard-curve model, matrix, sample preparation and preliminary assessment of assay robustness. The robustness of an assay is determined by its consistency under real life changes. Changes that have an impact on the assay must be identified. Elements in an immunoassay that could impact its consistency include incubation temperatures, incubation time, incubation with or without shaking, light exposure, change of assay reagent batches and lots, changes of stop solution volume and stopping time, coating time (e.g. over week-end), sample volume, or other factors. It is recommended to assess these factors during the assay development.

The critical assay reagents identified and optimized during the development phase should be used during the feasibility study and validation with no further change.

3. Feasibility study: milestone 1

Before any validation, a feasibility study should be performed to assess the performance of the available kit(s) or the home-made immunoassay. At the end of the feasibility study, a decision will be taken to either continue with the validation or change the assay.

A bioanalytical study plan describing the appropriate procedure as well as the parameters and their respective acceptance criteria should be written and signed before the feasibility study starts.

The feasibility study addresses the likelihood that an assay will be able to achieve its intended purpose. Parallelism, selectivity and reproducibility should be assessed using biological sample which should mimic the clinical population. Incurred samples should be used whenever possible to ensure that the assay is able to quantify the selected biomarker in the clinical population. When patient/target population samples can be made available, comparison between healthy population and patients should be made. If patient samples are not available during validation, it is recommended to randomly select samples retroactively to characterize the variability in the diseased population. Samples should not be pooled unless volumes are too small (in the case of rare matrices such as cerebrospinal fluid or synovial fluid). QCs are not used during the feasibility study. Therefore, the analytical runs are validated based on calibration curve acceptance criteria.

The result of the feasibility study will be incorporated in the validation report, and criteria evaluated will be used for the validation.

During the feasibility study, the parameters tested are identical whether the validation applies to decision-making biomarker or biomarker selection. The accuracy and precision acceptance criteria are summarized in Table 1 and are further detailed in the corresponding paragraph below. **Throughout the document, precision (CV) is calculated based on the concentration and not on the raw signal**. The CV on the concentration reflects the working range of the assay. The typical concentration–response relationship generated for an immunoassay is not linear. The accuracy and precision of the back-calculated STDs concentration will increase at the asymptote at the low and high values. Therefore, the CV on concentration is a useful tool during the assay development and optimization because it affords to estimate the lower and upper limit of quantification.

3.1. Calibration curve

STDs are prepared by spiking known amounts of the reference material into the appropriate buffer or surrogate matrix (such as a matrix depleted of the endogenous biomarker to be quantified). The protein-containing buffer or "substituted matrix" is our preferred option as it offers several advantages as opposed to the surrogate matrix: better stability, robustness, reproducibility. STDs must be freshly prepared and a calibration curve must be carried out for each run.

The calibration curve should consist of at least six non-zero STDs (excluding blank and anchor points) prepared in buffer (when less than six non-zero STDs are included in a kit, additional ones will be prepared by diluting or mixing already existing STDs (whenever possible)). The calibration standards should be spaced evenly within the anticipated working range on a logarithmic scale. For the curve within a run, the bias and precision of back-calculated value for at least 75% of the calibration standards should lie within 20% [10,20].

A concentration–response relationship is generated under established assay conditions, from which the biomarker concentration in unknown samples is interpolated. Both linear and well-characterized non-linear calibration relationships are acceptable; however, the preferred regression for a typical immunoassay is the 4/5-parameter logistic function. Selection of the regression model should have been rationalized during the method development. A STD can be excluded (or masked) from the calibration curve if it does not change the regression model used.

Number of STDs included in the curve	6	7	8
Acceptable number of excluded STD(s)	0	1	2

If it is necessary to deviate from the above-mentioned criteria for a particular assay, the reason should be documented in the raw data and in the validation report.

3.2. Parallelism

During the feasibility study, neat (non-spiked) biological samples will be tested pure and after dilution (direct or serial) to assess the endogenous value of the biomarker and to evaluate the parallelism of the assay. Samples will be diluted in assay buffer (the same as that used to prepare the STDs).

A minimum of six individual matrices should be tested during the feasibility study (one determination performed per matrix and per dilution factor).

The run will be validated based on calibration curve acceptance criteria.

As mentioned in the introduction parallelism assessment is very often missing from the publication for immunoassay validation or mistaken with selectivity [24–28]. In previous publications, authors recommend to verify parallelism by either performing a linear regression analysis of the plotted measured concentrations of the

Table 1
Summary of the parameters and their respective accuracy and precision acceptance criteria assessed during the feasibility study.

Parameters	Acceptance criteria	
	Biomarker selection	Decision-making biomarker
Calibration curve	≥6 non-zero STDs excluding blank and anchor point(s) Mean ACC: 80–120% CV: ≤20% ≥75% of non-zero STDs should meet the acceptance criteria	
Parallelism	6 individual neat matrices	Mean ACC: 70–130% compared to neat matrix undiluted or at the minimum required dilution and CV ≤ 25% For each dilution factor, at least 5/6 matrices should meet the acceptance criteria. If not, 6 additional matrices should be tested
Selectivity	6 individual matrices freshly spiked	Mean ACC: 70–130% compared to freshly spiked matrix undiluted or at the minimum required dilution and CV ≤ 25% For each dilution factor, at least 5/6 matrices should meet the acceptance criteria. If not, 6 additional matrices should be tested
	Decision point	

diluted samples against the 1/dilution factor using log scales or determining the CV amongst the measured concentrations at different dilutions. For the first approach, parallelism is proven when results show a slope of nearly 1 but acceptance criteria on the slope are not clearly defined [14,29]. For the second approach, parallelism is proven when CV amongst results at different dilutions is below 30% [17]. None of the two approaches allows the identification of a minimum dilution which is required to reach acceptable accuracy and precision. That is the reason why we recommend assessing parallelism by determining the accuracy and the precision of the measured concentration in each diluted samples. In a first attempt, the neat (undiluted) sample is used as "nominal" value to determine the accuracy. Parallelism is proven when accuracy and precision are within the defined acceptance criteria. When acceptance criteria are not met, it is very often the sign that a minimum dilution is required. In that case, in a second attempt, the accuracy should be calculated using the diluted sample (at the minimum required dilution (MRD)) as the nominal value. The CV on concentration is also a useful tool to estimate the lower limit of quantification as the precision is expected to decrease when the measured concentration reach the lower limit of quantification.

The acceptance criteria selected to prove parallelism are aligned with the acceptance criteria selected to assess the assay performance (the intra- and inter-run accuracy and precision) (see Table 3).

As a conclusion, parallelism is a key assessment as it serves several purposes: the assessment of the endogenous level of the biomarker in the biological sample; the demonstration of the proportionality between the endogenous form of the biomarker and the reference standard; the identification of minimum magnitude of dilution required to achieve acceptable accuracy and precision; and the estimation of the lower limit of quantification in the biological samples.

3.3. Selectivity

Selectivity is assessed by quantifying the biomarker in presence of endogenous matrix components that could potentially interfere with the antibodies binding.

The same six individual matrices used to assess the parallelism should be spiked with a known amount of reference material and analyzed after dilution (direct or serial) in assay buffer (the same as that used for the STDs).

The level of reference material to be spiked will be based on the endogenous biomarker level measured during the parallelism experiment. If endogenous level is close to the expected LLOQ, it may be needed to spike different levels of reference material: one spike close to the endogenous level and a second spike made at the expected ULOQ (when this is not feasible, efforts should be made to spike a concentration as high as possible). If endogenous level covers the whole working range, one level of spike should be enough. A highly concentrated stock solution of recombinant protein is often required as the spiking solution should represent less than 5% of the final volume [18]. In practice, the recombinant protein used to spike the matrix may differ from the recombinant protein used to prepare the calibration standard (different expression cell lines, buffer, concentration. . .). This is the reason why we do not recommend defining acceptance criteria for the recovery of the spiked reference material. The recovery is deemed acceptable as long as the recovery yield is constant for each dilution factor tested and for each matrix tested [23]. Recovery should be assessed taking into account the endogenous level when it is above the expected LLOQ.

In a first attempt, the spiked undiluted sample is used as "nominal" value to determine the accuracy. When a quantifiable amount of biomarker is present in the matrix, the use of an assigned value (measured concentration) is recommended as the recovery from the endogenous biomarker and added reference standard may not always behave in an additive manner (linear) [10]. Selectivity is proven when accuracy and precision are within the defined acceptance criteria. When acceptance criteria are not met, it is very often the sign that a minimum dilution is required. In that case, the accuracy should be calculated using the spiked diluted sample (at MRD) as the nominal value. The acceptance criteria required to prove selectivity are aligned with the acceptance criteria selected to assess the intra- and inter-run accuracy and precision (see Table 3).

The run will be validated based on calibration curve acceptance criteria.

4. Decision point: milestone 2

If the performance of the immunoassay obtained during the feasibility study is not satisfactory in terms of parallelism, the decision should be taken to evaluate a new kit or develop a new assay (using new antibodies or reagents). If the immunoassay lacks sensitivity, the decision may be taken to proceed with the validation if justified by a strong scientific and biological rationale (for example, a higher endogenous level of biomarker is expected in treated- or disease-sample).

The aim of milestone 2 is to avoid validating an immunoassay which is unlikely to meet the intended use. By taking an early decision, we limit the time and cost associated with the immunoassay validation.

Based on the result of parallelism and selectivity experiment, a MRD may be defined. It corresponds to the minimum magnitude of dilution required to achieve acceptable accuracy and precision in the quantification of both neat (endogenous) and spiked biological samples. There may be cases where the MRD identified during selectivity is higher than the one defined during parallelism. The decision may be taken not to apply the MRD defined during selectivity if it jeopardizes the immunoassay sensitivity and its ability to quantify the endogenous biomarker in the population of interest.

5. Definition of quality controls nominal value: milestone 3

Before the validation starts, the nominal/theoretical value of each QC level will be established. The true nominal value of each QC level will be determined from multiple runs performed without defined accuracy and precision acceptance criteria, used to assess the assay variability.

The difference between the decision-making biomarker and biomarker selection validation lies in the number of QC levels needed: three levels for "biomarker selection" and five levels for "decision-making" biomarker.

The parameters and the run acceptance criteria are summarized in Table 2.

QCs used to validate an assay should be prepared using the biological matrix (ideally derived from the specific trial population). QCs are distinct from the STDs and should not be used as part of the calibration curve. Preferred suppliers and commercial vendors for access to these relevant matrices should be consulted. When the matrix is difficult to procure an attempt should be made to prepare at least one level of QC in the actual matrix. The source of the biological matrix and lot used are to be stated in the validation protocol and references to documentation describing the characteristics of the material need to be enclosed (i.e. species, gender, known disease, age, storage conditions, etc.). In case of plasma, the same type of anticoagulant has to be used over the whole validation. The anticoagulant should be the same used in the studies.

In the assay validation for biomarker selection, three levels of QC are needed: QC LLOQ, QC MID and QC ULOQ. In the assay validation for decision-making biomarker, five levels of QC are required: QC LLOQ, QC LOW (within three times the LLOQ), QC MID, QC HIGH (70–80% of ULOQ) and QC ULOQ. A minimum of two individual matrices should be used to prepare the different levels of QC. Based on the result of parallelism and selectivity experiment, the decision will be taken to prepare the different QC level using either only neat matrix (pure or diluted in assay buffer to reach the targeted LLOQ for example), only spiked matrix or a combination of spiked and neat samples. **The preference is given to neat matrix especially at expected LLOQ**. **If QC samples need to be spiked, the endogenous biomarker level should be taken into consideration for the selection of spike level**. The spiking solution should represent less than 5% of the final volume. The QCs should be diluted according to the MRD if one was defined.

The concentration of each QC will be measured in three independent runs, three independent preparations per QC level and per run. It is recommended to perform the three runs on at least 2 different days. For each QC, the mean of the three experiments will be determined and used as nominal value during the validation.

Each run will be validated based on calibration curve acceptance criteria.

6. Validation: milestone 4

Validation performed for decision-making biomarker or biomarker selection will differ in terms of parameters tested, number of QCs required and acceptance criteria applied. Based on the result of feasibility study, the acceptance criteria may be adapted if justified.

The differences are summarized in Table 3.

6.1. Calibration curve

Based on result of feasibility study, it may be necessary to modify the STDs at the expected LLOQ and ULOQ.

- The lowest STD should be below or at the concentration corresponding to the QC LLOQ
- The highest STD should be above or at the concentration corresponding to the QC ULOQ

STDs must not be changed once assay validation starts.

6.2. Intra- and inter-run precision and accuracy

The conduct of accuracy and precision experiments for biomarkers is similar to that for pharmacokinetic with 6 runs to determine inter-assay variability and 2–5 replicates of QCs in each run for determination of intra-assay variability. However, the number of QCs and assay runs may be less for exploratory method validation, as justified by the intended use of the method and following the validation plan [10,18,20]. This is reflected in the different number of QCs and validation runs recommended in the validation of decision-making biomarker and biomarker selection.

Recognized performance standards have been established for the definitive quantitative assays to support pharmacokinetic assessments where repeat analyses of the QCs are expected to vary <15% with 20% acceptable at the LLOQ for both accuracy and precision. More flexibility is however allowed in assay validation to support biomarker assessment because of the potential impact the endogenous biomarker may have on accuracy and precision [17,20]. It has been shown that accuracy and precision are increased when QCs are prepared in matrix compared to QCs prepared in buffer [19]. However accuracy and precision data of QCs prepared in the biological matrix are believed to represent more realistically the assay performance of clinical samples. This is especially true for the QC targeting the lower limit of quantification where the assay variability is increased [18,19]. We recommend using neat (non-spiked) QCs rather than spiked sample as they truly reflect the trial population. So relaxation of the acceptance criteria is justified by the potential impact the endogenous biomarker may have on the accuracy and precision of the measurement. During validation of decision-making biomarker, 25% is the default value (30% at the LLOQ and ULOQ) for precision and accuracy [17,20]. The relaxation of the acceptance criteria is more pronounced for the second level of validation (biomarker selection) when biomarkers are considered exploratory endpoints and used to support exploratory studies or objectives.

Each run should contain a calibration curve, a blank and QCs covering the anticipated working range. It is recommended to perform the validation runs on at least three different days. Each QC nominal value is previously defined in the feasibility study. Five independent determinations of each level of QC will be performed for each validation run.

The intra-run accuracy and precision will be calculated for each QC level as the mean accuracy and the precision of the five determinations for each validation run. The inter-run accuracy and precision will be calculated for each QC level as the mean accuracy and precision of all measurements performed. All results – other than those rejected for analytical/technical reasons (i.e. pipetting error) – will be used in the calculation of accuracy and precision. If the acceptance criteria are not met (e.g. showing a higher variability related to the compound, matrix or antigen–antibody reaction),

Table 2
Summary of the parameters and their respective accuracy and precision acceptance criteria assessed during the definition of QCs nominal value.

Parameters	Acceptance criteria	
	Biomarker selection	Decision-making biomarker
Calibration curve	≥6 non-zero STDs excluding blank and anchor point(s) Mean ACC: 80–120% CV: ≤20% ≥75% of non-zero STDs should meet the acceptance criteria STDs at the expected LLOQ and ULOQ should meet the acceptance criteria	
QC nominal value definition	3 QC level: LLOQ, MID, ULOQ 3 independent runs 3 independent preparations per QC level	5 QC level: LLOQ, LOW, MID, HIGH, ULOQ 3 independent runs 3 independent preparations per QC level

new acceptance criteria must be defined and documented accordingly before applying those to study sample measurements.

For a method to be considered acceptable, its performance should meet the criteria defined above in terms of accuracy and precision within each validation run. It is also recommended that the inter-assay accuracy and precision both be <20% (25% at LLOQ and ULOQ). However a method may be considered acceptable based on its performance but not adapted to its application in a clinical context. Other factors should be considered beyond the assay performance: the intended use of the data (objectives of the clinical study), biological variability, and expected effect on the biomarker of interest. These factors (when they are known) should be taken into consideration when defining the acceptance criteria for a biomarker assay validation. When the expected effect on the biomarker of interest is unknown, the performance of the biomarker assay defined during the validation should be taken into consideration when defining its intended application and predictive power.

6.3. Working range/sensitivity

The validation experiment establishes the working range defined by the QC LLOQ and the QC ULOQ, whereby the sensitivity of the method is based on the QC LLOQ, i.e. the lowest concentration of the biomarker in a biological sample that can be quantitatively determined with acceptable precision and accuracy.

6.4. Parallelism

Parallelism is already evaluated during the feasibility study using a minimum of six individual matrices. Six additional matrices should be tested during the validation for "decision-making" biomarker.

When the endogenous biomarker cannot be quantified in the "healthy controls" used during the validation, retrospective validation based on incurred samples is strongly recommended if sufficient amounts of the clinical samples remain following sample analysis.

Each run should contain a calibration curve, a blank and the five QC levels (two independent preparations per QC). The run will be validated based on calibration curve and QCs acceptance criteria.

6.5. Selectivity

Selectivity is already evaluated during the feasibility study using a minimum of six individual matrices spiked with a known amount of reference material. Six additional matrices should be tested during the validation for a "decision-making" biomarker.

Each run should contain a calibration curve, a blank and the five QC levels (two independent preparations per QC). The run will be validated based on calibration curve and QCs acceptance criteria.

6.6. Dilution linearity

If dilution of study samples is anticipated during sample analysis, linearity of dilution must be demonstrated. It is necessary to show that the biomarker of interest, when it is present in concentration above the ULOQ, can be diluted to bring the biomarker concentration into the working range of the assay. An additional reason for conducting dilution experiment is for the identification of a possible prozone or "hook effect".

Dilution linearity should be evaluated on freshly spiked samples. Dilutions should be made such that several dilutions fall into the working range. Evaluation of dilution linearity should be done with a spike made 100- to 1000-fold greater than the ULOQ. When this evaluation is not feasible, efforts should be made to prepare a spiked concentration as high as possible. The dilution samples that are evaluated should be designed to include a dilution just above the ULOQ (to evaluate the hook effect). Generally, it is common practice that an individual dilution step does not exceed 1:100.

A minimum of three individual matrices should be tested. Incurred samples should be used whenever possible.

Each run should contain a calibration curve, a blank and the five QC levels (two independent preparations per QC). The run will be validated based on calibration curve and QCs acceptance criteria.

6.7. Specificity

Specificity of the antibody(ies) should be tested during the assay validation of a "decision-making" biomarker. The antibody used should be specific for the biomarker of interest, without cross-reactivity with variant forms of the biomarker or other structurally related compounds that may be present in the sample.

At least three individual matrices should be spiked with a known amount of a recombinant protein which belongs to the same family and/or presents chemical similarities to the biomarker of interest and analyzed after dilution (direct or serial) in assay buffer. Two different levels of the analogue biomarker should be tested: one level which reflects its endogenous level and a second concentration which represents a 100-fold excess compared to the biomarker of interest. The recovery will be assessed.

Each run should contain a calibration curve, a blank and the five QC levels (two independent preparations per QC). The run will be validated based on calibration curve and QCs acceptance criteria.

6.8. Stability in biological matrix

Stability of the endogenous biomarker in the biological matrix and under specific assay condition must be evaluated.

Table 3
Summary of the parameters and their respective accuracy and precision acceptance criteria assessed during the validation.

	Biomarker selection		Decision-making biomarker	
	Parameters	Acceptance criteria	Parameters	Acceptance criteria
Calibration curve	≥ 6 non-zero STDs Mean ACC: 80–120% CV: ≤20% ≥75% of non-zero STDs should meet the acceptance criteria STDs at LLOQ and ULOQ should meet the acceptance criteria			
Intra-, inter-run precision and accuracy	3 QC levels	Mean ACC: 70–130%	5 QC levels	Mean ACC: 75–125%
	3 independent runs 5 independent preparations per QC level	CV ≤ 25% 4/5 determination per QC level should meet the acceptance criteria	6 independent runs 5 independent preparations per QC level	(70–130% at LLOQ and ULOQ) CV ≤ 25% 4/5 determination per QC level should meet the acceptance criteria
Parallelism	No further assessment		6 additional individual neat matrices	Mean ACC: 75–125% compared to undiluted neat matrix or matrix at MRD (70–130% at LLOQ and ULOQ) CV ≤ 25% For each dilution factor, at least 5/6 matrices should meet the acceptance criteria. If not, 6 additional matrices should be tested
Selectivity	No further assessment		6 additional individual matrices freshly spiked	Mean ACC: 75–125% compared to undiluted spiked matrix or spiked matrix at MRD (70–130% at LLOQ and ULOQ) CV ≤ 25% For each dilution factor, at least 5/6 matrices should meet the acceptance criteria. If not, 6 additional matrices should be tested
Dilution linearity	Not assessed (unless biological evidence)		Minimum 3 individual matrices freshly spiked	Mean recovery: 75–125% (70–130% at LLOQ and ULOQ) CV ≤ 25% 2/3 matrices should meet the acceptance criteria
Antibody specificity	Not assessed		Minimum 3 individual matrices	Mean recovery: 75–125% (70–130% at LLOQ and ULOQ) CV ≤ 25% 2/3 matrices should meet the acceptance criteria
Short-term and bench top stability	2 individual neat matrices	Mean recovery: 70–130%	2 individual neat matrices	Mean recovery: 75–125%
	3 independent preparations per matrix	CV ≤ 25% At least 4/6 matrices should meet the acceptance criteria	3 independent preparations per matrix	(70–130% at LLOQ and ULOQ) CV ≤ 25% At least 4/6 matrices should meet the acceptance criteria
Long-term stability	Not assessed		2 individual matrices 3 independent preparations per matrix	Mean recovery: 75–125% (70–130% at LLOQ and ULOQ) CV ≤ 25% At least 4/6 matrices should meet the acceptance criteria

To assess the stability of the biomarker of interest, two individual matrices will be used. **Neat matrices should preferentially be used even if they do not cover the working range**. However, if it is not possible, freshly spiked sample can be used. In that case, the spiked stability samples should cover the low and high level of the working range. The stability samples should be stored undiluted. When the validation is performed using only spiked QCs, it is strongly recommended to use clinical samples to assess the stability of the endogenous biomarker (if sufficient amounts of the clinical samples remain following sample analysis).

Three independent preparations of each matrix should be analyzed. The mean concentration measured after a defined storage period/condition is compared to the concentration measured at day 0. Day 0 is performed using stability samples which have undergone one freeze/thaw cycle and before storage is initiated.

Short-term stability in biological matrix will be assessed at room temperature, 4 °C, −20 °C and −80 °C for the time required to prepare samples and complete the assay (if it is not possible to perform the two experiments (day 0 and day 0–4 h) on the same day, then stability will be tested after 24 h) and 2 weeks. Short-term stability will also be assessed at 4 °C, −20 °C and −80 °C for 1 month. It is critical to test the short-term stability at 4 °C and −20 °C as many clinical sites do not have −80 °C freezers or have a limited storage capacity.

The long-term stability in biological matrix will only be assessed for "decision-making" biomarker at −20 °C and −80 °C for 3, 6, 12 and 24 months. Long-term stability should be demonstrated over a period of time, which exceeds the time between the collection and analysis of any given study sample.

Each run should contain a calibration curve, a blank and three or five QC levels (two independent preparations per QC) depending on the validation procedure. The run will be validated based on calibration curve and QCs acceptance criteria.

6.9. Freeze and thaw stability

Stability samples (the ones described above to assess the short-term and long-term stability) will be subjected to a minimum of five 'freeze-thaw' cycles. Samples should be stored at −20 °C and −80 °C for at least 24 h and thawed unassisted at room temperature. When completely thawed, the samples should be left at room temperature or on ice (depending on the assay protocol) for the time required to prepare samples, then transferred back to the original freezer and kept frozen for at least 24 h. This cycle of thawing and freezing should be repeated four more times, and then the stability samples are analyzed on the fifth cycle.

When a biomarker is not stable after five freeze and thaw cycles, the stability should be tested after only four freeze and thaw cycles. In case of instability, three freeze and thaw cycle should be tested and so on.

Each run should contain a calibration curve, a blank and three or five QC levels (two independent preparations per QC) depending on the validation procedure. The run will be validated based on calibration curve and QCs acceptance criteria.

6.10. Assay robustness

The robustness of an assay is determined by its consistency under real life changes. Changes that have an impact on the assay must be identified in the method. Elements in an immunoassay that could impact its consistency include incubation temperature, incubation time, incubation with or without shaking, light exposure, change of assay reagent batches and lots, changes of stop solution volume and stopping time, coating time (e.g. over weekend), sample volume, or other factors. It is recommended to assess these factors during the assay development. The exact assay procedure and critical assay conditions, if identified, are to be mentioned in the assay documentation.

When changes are made to a previously validated assay, the analyst should exercise good judgment as to how much additional validation is needed. For changes considered critical in nature, and this may include sample-processing changes, extension of concentration range, conversion of manual to automatic method, changes of instrument, or qualification of an analyst, a cross-validation is recommended. Changes of kit or reference material lot may only require one run to confirm equivalent performance. The run should contain a calibration curve, a zero STD and three or five QC levels (five independent preparations per QC) depending on the validation procedure. The run will be validated based on the QC acceptance criteria.

7. Assay requalification: milestone 5

When the assay has not been performed for six months, the assay performance must be demonstrated by analyzing at least a calibration curve and three or five levels of QC (five independent preparations for each QC level) according to the validation performed. If stability is demonstrated over 6 months, QCs used during the validation may be used to test the assay performance. If it is not possible, new QCs should be prepared and the nominal value should be defined (as described during the feasibility study).

8. Conclusion

This paper describes the implementation of a bioanalytical study plan for the validation of immunoassays to support decision-making biomarkers and biomarker selection during preclinical and clinical studies. It establishes clear operating procedures as well as the parameters and their respective acceptance criteria and proposes milestones and decision points to be followed during the assay validation that should result in high quality bioanalytical data in a limited timeframe and with reduced costs.

Amongst these milestones, parallelism is the most critical one. It is recognized that the extent of the validation should be correlated to the intended use of the data and the assay acceptance criteria should take into consideration the study objective(s), nature of the methodology and the biological variability of the biomarker. The bioanalytical study plan has been successfully applied to the validation of immunoassay based on classical monoplex ELISA, automated analyzer (Elecsys platform), multiplex planar array platform (MesoScale Discovery), multiplex suspension microbead platform (Luminex) and the flow-based single-molecule counting Erenna® platform (Singulex).

Acknowledgements

We would like to thank Dr Joanne Meyer and Dr Julie Doucet from Novartis Institute for Biomedical Research for critically reviewing the manuscript.

References

[1] Food and Drug Administration, Draft Guidance for Industry: Bioanalytical Method Validation, Food and Drug Administration, Rockville, MD, 1999.
[2] V.P. Shah, The history of bioanalytical method validation and regulation: evolution of a guidance document on bioanalytical methods validation, AAPS J. 9 (2007) E43–E47.
[3] Guidance for Industry on bioanalytical method validation: availability, Fed. Reg. 66 (2001) 28526–28527.
[4] EURACHEM Working Group, The Fitness for Purpose of Analytical Methods: A Laboratory Guide to Method Validation and Related Topics, 1998.
[5] ICH Topic Q2 Validation of Analytical Methods, The European Agency for the Evaluation of Medicinal Products. ICH Topic Q 2 A – Definitions and Terminology. ICH Topic Q 2 B – Methodology.

[6] Food and Drug Administration, Draft Guidance for Industry: Analytical Procedures and Methods Validation, Food and Drug Administration, Rockville, MD, 2000.
[7] European Medicines Agency, Concept paper/recommendations on the need for a (CHMP) guideline on the validation of bioanalytical methods, EMEA/CHMP/EWP/531305, 2008.
[8] European Medicines Agency, Draft Bioanalytical Methods, EMEA/CHMP/EWP/192217, 2009.
[9] C.T. Viswanathan, S. Bansal, B. Booth, et al., Quantitative bioanalytical methods validation and implementation: best practices for chromatographic and ligand binding assays, Pharm. Res. 24 (2007) 1962–1973.
[10] B. DeSilva, W. Smith, R. Reiner, et al., Recommendations for the bioanalytical method validation of ligand-binding assays to support pharmacokinetic assessments of macromolecules, Pharm. Res. 20 (2003) 1885–1900.
[11] G. Smith, Bioanalytical method validation: notable points in the 2009 draft EMA Guideline and differences with the 2001 FDA Guidance, Bioanalysis 2 (2010) 929–935.
[12] Code of Federal Regulations, Clinical Laboratory Improvement Amendment, Title 42, 2001.
[13] J.W. Lee, R.S. Weiner, J.M. Sailstad, et al., Method validation and measurement of biomarkers in nonclinical and clinical samples in drug development: a conference report, Pharm. Res. 22 (2005) 499–511.
[14] J.W. Lee, V. Devanarayan, Y.C. Barrett, et al., Fit-for-purpose method development and validation for successful biomarker measurement, Pharm. Res. 23 (2006) 312–328.
[15] J.W. Lee, Method validation and application of protein biomarkers: basic similarities and differences from biotherapeutics, Bioanalysis 1 (2009) 1461–1474.
[16] C.H. Chau, O. Rixe, H. McLeod, et al., Validation of analytical methods for biomarkers employed in drug development, Clin. Cancer Res. 14 (2008) 5967–5976.
[17] J. Cummings, F. Raynaud, L. Jones, et al., Fit-for-purpose biomarker method validation for application in clinical trials of anticancer drugs, Br. J. Cancer 103 (2010) 1313–1317.
[18] J.W. Lee, M. Hall, Method validation of protein biomarkers in support of drug development or clinical diagnosis/prognosis, J. Chromatogr. B 877 (2009) 1259–1271.
[19] J. Wang, J. Lee, D. Burns, et al., "Fit-for-purpose" method validation and application of a biomarker (C-terminal telopeptides of type 1 collagen) in Denomusab clinical studies, AAPS J. 11 (2009) 385–394.
[20] M. Kelley, B. DeSilva, Key elements of bioanalytical method validation for macromolecules, AAPS J. 9 (2007) E156–E163.
[21] S. Bansal, A. DeStefano, Key elements of bioanalytical method validation for small molecules, AAPS J. 9 (2007) E109–E114.
[22] C. Brasse-Lagnel, M. Poli, C. Lesueur, et al., Immunoassay for human hemojuvelin, Haematologica 95 (2010) 2031–2037.
[23] R. Houghton, C. Horro Pita, I. Ward, et al., Generic approach to validation of small-molecule LC–MS/MS biomarker assays, Bioanalysis 1 (2009) 1365–1374.
[24] K. St. Ledger, S.J. Agee, M.T. Kasaian, et al., Analytical validation of a highly sensitive microparticle-based immunoassay for the quantitation of IL-13 in human serum using the Erenna immunoassay system, J. Immunol. Methods 350 (2009) 161–170.
[25] S. Schellenberg, B. Grenacher, K. Kaufmann, et al., Analytical validation of commercial immunoassays for the measurement of cardiovascular peptides in the dog, Vet. J. 178 (2008) 85–90.
[26] C.A. Ray, R.R. Bowsher, W.C. Smith, et al., Development, validation, and implementation of a multiplex immunoassay for the simultaneous determination of five cytokines in human serum, J. Pharmaceut. Biomed. Anal. 36 (2005) 1037–1044.
[27] A.A. Ellington, I.J. Kullo, K.R. Bailey, et al., Antibody-based protein multiplex platforms: technical and operational challenges, Clin. Chem. 56 (2010) 186–193.
[28] Q. Fu, J. Zhu, J.E. Van Eyk, Comparison of multiplex immunoassay platforms, Clin. Chem. 56 (2010) 314–318.
[29] J.W.A. Findlay, W.C. Smith, J.W. Lee, et al., Validation of immunoassays for bioanalysis: a pharmaceutical industry perspective, J. Pharmaceut. Biomed. Anal. 21 (2000) 1249–1273.
[30] Biomarkers Definitions Working Group, Biomarkers and surrogate endpoints: preferred definitions and conceptual framework, Clin. Pharmacol. Ther. 69 (2001) 89–95.

Glossary

A *biomarker* is "a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic response to a therapeutic intervention" [30]

A *clinical end-point* is defined as a "characteristic that reflects how a patient feels, functions or survives" [30]

A *surrogate end-point* is a "biomarker intended to substitute for a clinical end-point" [30]

Analytical method *validation* is the process of assessing the performance characteristics of a given assay; whereas, clinical *qualification* is the evidentiary and statistical process linking biologic, pathologic and clinical endpoints to the drug effect or linking a biomarker to biologic and clinical endpoints. The term "validation" is recommended to be used for analytical methods [30]

R*eference material*: used to prepare STDs and spiked QCs. The reference material, if possible, should be identical to the endogenous biomarker. When this is not possible, an established and well characterized recombinant form can be used. Three types of reference material are usually used: (1) certified reference material (International WHO or NIBSC standards), (2) commercially available reference material obtained from a reliable source and/or (3) material with documented purity which is custom-synthesized by an analytical laboratory or other non-commercial establishment. The reference material should derive consistently from the same supplier. The comparability between the batches needs to be demonstrated, if different batches are used during validation. To avoid any stability issue, stock and working solutions of the reference material should be prepared in concentration as high as possible in the appropriate buffer *Accuracy (ACC):* Closeness of the measured concentration to the nominal or known true concentration of the biomarker ACC (%): (measured value/nominal value) × 100

Anchor point: Calibration standard outside of the working range which are recommended in order to facilitate the fitting of the standard curve. They are not required to meet the acceptance criteria

Blank: Zero STD prepared in buffer or surrogate matrix without reference material. It is not part of the calibration curve

Calibration curve: Numerical relationship between experimental response values and the analytical concentrations of the biomarker. Also defined as the standard curve

Calibration standard (STD): A sample to which a known amount of reference material has been added. STDs are used to construct calibration curves from which the biomarker concentrations in quality control and in unknown study samples are determined. STDs are prepared in buffer or surrogate matrix.

Determination: Analysis of one sample, which is usually measured in duplicate or triplicate wells from a single sample preparation. The result of one determination is the mean value of the replicate wells

Linearity of dilution: Ability of an analytical method to obtain test results that are directly proportional to the concentration of biomarker in the samples upon dilution of high concentration samples (at or above the ULOQ) through the quantitative range of the assay

Lower limit of quantification

(LLOQ): The lowest concentration of biomarker in a biological sample that can be quantitatively determined with acceptable precision and accuracy

Matrix: Any biological specimen (i.e. plasma, serum, urine, cerebrospinal fluid) in which the biomarker is quantified

Parallelism: Assessment of dilutional linearity of neat (non-spiked) authentic biological sample. It demonstrates that the relationship between concentration and response for the biomarker in sample matrix is sufficiently similar to the reference material being measured in substitute matrix

Precision (CV): Closeness of individual concentrations of a biomarker when the method is applied repeatedly to multiple aliquots of the same biological sample CV (%): (Standard Deviation/mean concentration) × 100

Quality control (QC): A neat or spiked biological sample used to monitor the performance of a bioanalytical method, adopted to assess the integrity and validity of the results of the unknown samples analyzed in an individual batch

Recovery: The quantified closeness of an observed result to its theoretical true value, expressed as a percent of the nominal (theoretical) concentration. Recovery is used to assess the amount of biomarker spiked in and recovered from a biological sample, compared to the neat (non-spiked) sample. Recovery (%): measured concentration/(neat concentration + nominal concentration) × 100

Reproducibility: Represents the precision of a method under the same operating conditions over a short period of time, also between two laboratories

Robustness: A measure of the capacity of an analytical procedure to remain unaffected by minor, but deliberated variations in method parameters

Selectivity: Ability of an analytical method to determine the biomarker unequivocally in the presence of components that may be in the biological sample (i.e. other biomarkers, concomitantly administered drugs, metabolites)

Sensitivity: Defined by the LLOQ

Specificity: Ability of an antibody used in analytical method to bind solely the biomarker of interest. Ideally, antibody used would be specific for the biomarker, without cross-reactivity with variant forms of the biomarker of interest or other structurally related compounds that may be present in the biological sample

Stability: The physico-chemical stability of a biomarker in a given solution or matrix under specific conditions for given time intervals

Surrogate matrix: Biomarker-free substitute matrix, which can be used as STDs diluents if endogenous level of the biomarker is present in the intended sample matrix

Upper limit of quantification

(ULOQ): The highest amount of a biomarker in a sample that can be quantitatively determined with suitable precision and accuracy

Working range: Range of biomarker concentrations defined between the LLOQ and the ULOQ