ELSEVIER

The review presents strategies for fit-for-purpose method validation of biomarker assays to help ensure generation of robust data during clinical trials and to satisfy regulatory requirements.

Fit-for-purpose biomarker method validation in anticancer drug development

Jeffrey Cummings, Tim H. Ward and Caroline Dive

Clinical and Experimental Pharmacology, Paterson Institute for Cancer Research, University of Manchester, Wilmslow Road, Manchester M20 4BX, England, United Kingdom

The introduction of new anticancer drugs into the clinic is often hampered by a lack of qualified biomarkers. Method validation is indispensable to successful biomarker qualification and is also a regulatory requirement. Recently, the fit-for-purpose approach has been developed to promote flexible yet rigorous biomarker method validation, although its full implications are often overlooked. This review aims to clarify many of the scientific and regulatory issues surrounding biomarker method validation and the analysis of samples collected from clinical trial subjects. It also strives to provide clear guidance on validation strategies for each of the five categories that define the majority of biomarker assays, citing specific examples.

Biomarkers continue to offer considerable potential to enhance the progress of clinical research and accelerate the pace of new drug development [1,2]. Nowhere is this more urgently required than in anticancer drug development, where traditionally the rate of compound attrition is high and success in the clinic low [3,4]. During clinical trials of anticancer drugs, predictive biomarkers might facilitate the selection of patients (enrichment) most likely to respond to molecularly targeted agents, whereas pharmacological biomarkers might enable real-time monitoring of drug action, treatment efficacy or early signs of toxicity [5,6]. Qualification, the evidentiary process of proving a linkage between the biomarker and a clinical endpoint, can take many years, requiring not only retrospective and prospective clinical trials but also large population screening, without any guarantee of eventual success [7,8]. Thus, in many modalities of cancer therapy, there remain lamentably few, if any, qualified biomarkers for patient selection and pharmacological evaluation of new agents in clinical trials [9]. Method validation remains a key determinant in the successful qualification of a biomarker, and often, the failure of a biomarker in the clinic is not due to the underlying scientific rationale but to a poor choice of assay and lack of validation [5,6].

Clinical trial regulations in the UK and in Europe state that 'systems with procedures that assure the quality of every aspect of the trial should be implemented', which includes method validation [10]. These regulations, however, make only vague references to the laboratories conducting trial sample analysis. More recently, both the MHRA in the UK and the EU have published guidance

DR JEFF CUMMINGS

BSc (Hons) PhD is a staff scientist working within the Clinical and Experimental Pharmacology Group of the Paterson Institute for Cancer Research, Manchester UK and is responsible for the implementation of the group's quality assurance system. He has over 28 years experience in cancer pharmacology and anticancer drug discovery and has been at the forefront of all the major developments in academic quality assurance in the UK over the past ten years. His present research interest is focussed on the application of bioinformatics to biomarker method validation and qualification during early phase clinical trials of anticancer drugs. He recently guest edited a special issue of the Journal of Chromatography on Quantitative Analysis of Biomarkers by LC/MS.

DR TIM WARD

BTech (Hons) PhD is a translational scientist working within the Clinical and Experimental Pharmacology Group of the Paterson Institute for Cancer Research, Manchester, UK. He is the pharmacodynamics manager responsible for the development, validation and implementation of all assays used by the group in early phase clinical trials. He has over 30 years experience in Cancer Research and Anticancer Drug Development specializing in anti-cancer drug screening and DNA damaging assays. He currently sits on the Cancer Research UK New Agents committee (NAC). His current research interests are centred on detecting and characterizing circulating tumor cells and micro-emboli. He also has a keen interest in utilizing circulating DNA as a biomarker of tumor DNA to detect specific tumor mutations which impact on response to modern targeted agents.

PROFESSOR CAROLINE DIVE

BPharm (Hons) PhD is the head of the Clinical and Experimental Pharmacology Group of the Paterson Institute for Cancer Research, Manchester, UK. Although maintaining a long standing interest in the basic regulation of drug-induced apoptosis, her focus has become progressively more translational, concentrating on validation and qualification of pharmacodynamic, predictive and safety biomarkers to facilitate drug development and personalized medicine for cancer treatment. Under her leadership, the group has developed close integration with the Early Clinical Trials Unit at the adjacent Christie Hospital, which is currently being enlarged and developed to become one of the largest of its kind worldwide. Her current research interests include evaluating biomarkers of tumor cell death, working with clinical colleagues on several apoptosis targeted novel agents entering Phase I/II Clinical Trials and exploring the biomarker utility and molecular characteristics of circulating tumour cells in lung cancer patients.

Corresponding author:. Cummings, J. (jcummings@picr.man.ac.uk)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2010.07.006

documents to aid laboratories in maintaining regulatory compliance to the clinical trials regulations, but again these lack specific recommendations for method validation (http://www.mhra.gov.uk/Howweregulate/Medicines/Inspectionandstandards/GoodClinicalPracticeforClinicalLaboratories/CON041197).

In the USA, the FDA also does not specify requirements for method validation but nonetheless provides comprehensive technique-based guidance documents [11]. The FDA CLIA regulations require accreditation for laboratories performing clinical testing and provide general guidelines for method validation, especially in the case of laboratory-developed tests [12]. Cancer Research UK places method validation at the heart of its roadmaps for qualification of biomarkers.

With an array of regulatory requirements and guidance documents, and especially considering that biomarker assays span a broad spectrum of technologies, it is perhaps not surprising that there is still a need for amplification on many of the issues associated with biomarker analysis [10,13]. A systematic approach that clearly distinguishes between the highly proscriptive 'good laboratory practice' method validation approach developed for small molecule bioanalysis by the pharmaceutical industry [14–17], has recently emerged and is referred to as 'fit-for-purpose' biomarker method validation [18,19]. This review aims to clarify the fit-for-purpose approach to biomarker method validation.

What is fit-for-purpose biomarker method validation?

Understanding fit-for-purpose biomarker method validation is founded on a proper and clear appreciation of the definition of the term 'method validation'. The benchmark definition for analytical method validation has been provided by the International Organisation for Standardisation as follows: 'the confirmation by examination and the provision of objective evidence that the particular requirements for a specific intended use are fulfilled' [20,21]. Although developed predominately for chemical analysis laboratories, it has been adopted – to a greater or lesser extent – by the bioanalytical field and the biomarker research community [12,15,18]. While this definition seems to be self-evident, its full implications are often overlooked [22–24]. Accordingly, method validation should proceed down two parallel tracks, which eventually converge – one experimental, the other operational (Fig. 1). The first track establishes the expectations of the sponsor or the scientific goals of the study based on existing scientific literature, then defines the role of the biomarker measurements in the clinical trial or investigation and, eventually, agrees upon outcomes, target values or acceptance limits. In parallel, the performance of the biomarker assay is characterized by experimentation, based on a previously agreed validation plan. The key stage in the whole process is the cross-comparison of the two strands leading to crucial evaluation of the technical performance of the assay against the predefined purpose (Fig. 1). If the assay with its newly established performance criteria can deliver to expectations, it is deemed fit for that purpose and valid. If not, then it cannot be deemed either fit for the specified purpose or valid.

Thus, comparing technical performance of an assay against predefined acceptance limits in isolation of purpose does not conform to the strict definition of method validation. Nonetheless,

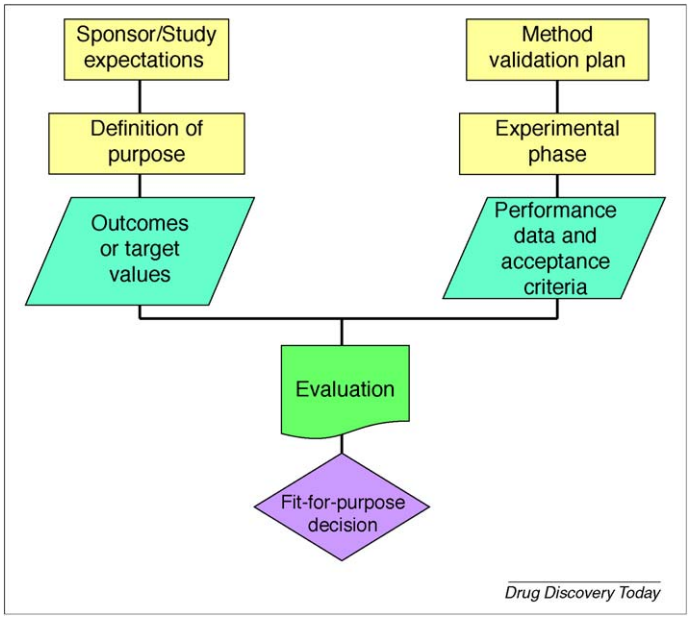

FIGURE 1

Fit-for-purpose biomarker method validation flow chart. Ideally, fit-for-purpose biomarker method validation should proceed down two parallel tracks, which eventually converge. The first is to establish the expectations of the sponsor or the scientific goals of the study to define the purpose of the assay in terms of outcomes, target values or acceptance limits. In parallel, the technical performance of the biomarker assay is characterized by experimentation. The key stage is the evaluation of whether the technical performance of the assay can deliver the predefined purpose. Based on Ref. [26], reproduced with permission from Elsevier.

performance data can play an important part in verifying that the assay is working properly and in aiding in the diagnosis of faults [25]. To quote a recent paper: 'validation determines that we are doing the *correct* test; verification confirms that we are doing the test *correctly*' [12].

Reliance on performance data alone can result in an anomalous situation in which an assay with tight performance would be rejected in favour of one with less tight performance [26] (Fig. 2). In the example given in Fig. 2, two different quantitative assays (such as LC/MS or ELISA) termed A and B both exhibit the same level of bias (systematic error resulting in consistent underestimation or overestimation compared to true values), but A has much tighter precision (random error) than B. On the basis of performance alone, A consistently misses the target, whereas B does not. In fit-for-purpose validation, however, a ring of expected values (note, not a performance acceptance limit) set in advance is used to evaluate performance, and in this scenario A is clearly superior to B.

The fit-for-purpose definition of method validation and the flexibility this brings is particularly well suited to biomarkers because, by their nature, they can have many different purposes in a variety of contexts. During anticancer drug development, biomarkers are used as discovery tools in compound selection, as pharmacodynamic markers of drug mechanism or efficacy in early-phase trials, or as predictive indices of patient response in late-phase trials [18,27]. In experimental cancer medicine, biomarkers might provide a diagnostic readout on tumour biology or be prognostic of disease or therapeutic outcome [28]. In fit-for-purpose, the position of the biomarker in the spectrum between research tool and clinical endpoint dictates the burden

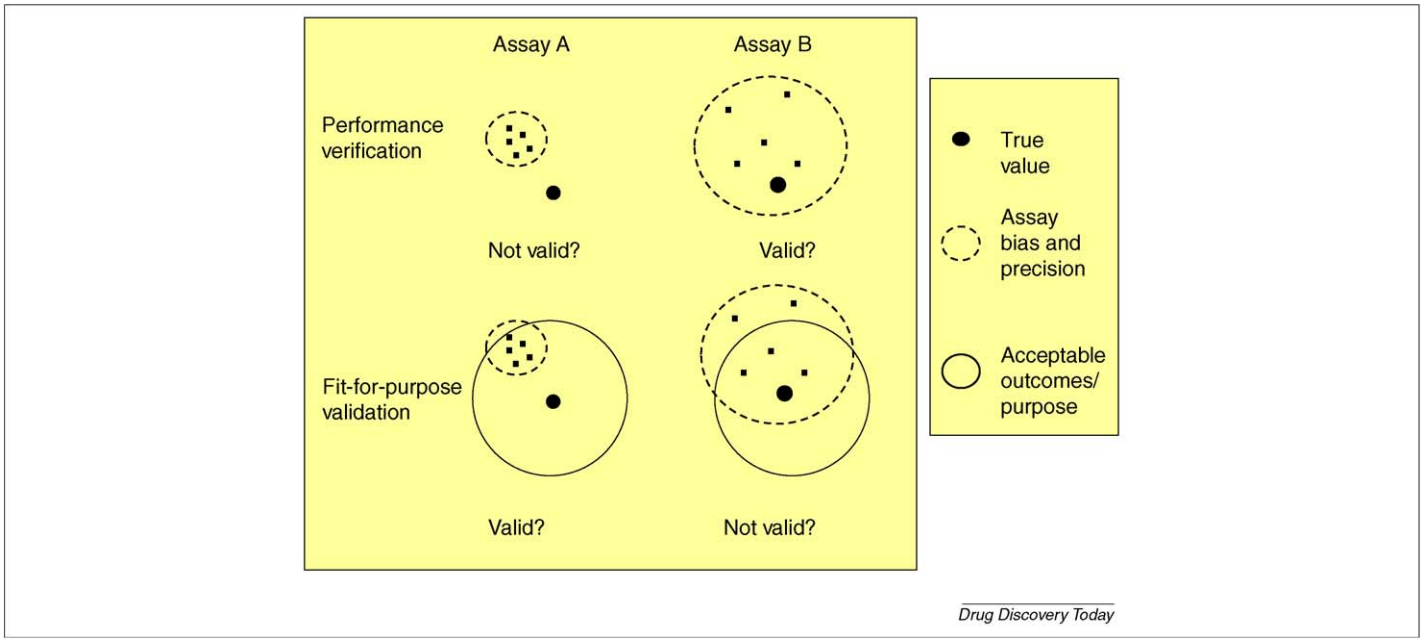

FIGURE 2

Performance verification versus fit-for-purpose validation. Schematic representation of the goals of classical validation with the emphasis on characterization of performance compared to fit-for-purpose validation with the emphasis on satisfying purpose. In classical validation, assay A exhibits better performance but consistently misses the target and, therefore, seems inferior to B. However, in fit-for-purpose validation, where the purpose is defined as a ring of expected values, assay A is clearly superior to B.
Based on Ref. [26], reproduced with permission from Elsevier.

of experimental proof required to achieve method validation [29]. The intrinsic nature of the analytical technology also greatly influences the level of performance verification required in validation strategies. For a quantitative diagnostic test, method validation would require that the assay is demonstrated to achieve an acceptable level of diagnostic accuracy because that is its purpose. Parameters such as analytic and clinical specificity (the ability to obtain negative analytical results in concordance with a negative confirmed clinical diagnosis) and analytical and clinical sensitivity (the ability to obtain positive analytical results in concordance with a positive confirmed clinical diagnosis) would have to be fully characterized [12]. Method validation might even include the use of Receiver Operating Characteristic curves and cut-off values to confirm purpose. By contrast, a putative biomarker used during discovery and measured by a commercially available ELISA might require no more than three assays to pass manufactures' acceptance criteria to verify performance, with no expectations placed on desired outcomes [19].

How to conduct fit-for-purpose biomarker method validation

In October 2003, a workshop cosponsored by the American Association of Pharmaceutical Scientists (AAPS) and the US Clinical Ligand Society was held in Salt Lake City, Utah, to address the unresolved issue of validation of biomarker assays in support of drug development [18]. At this workshop, five general classes of biomarker assays were identified [18]. A definitive quantitative assay makes use of calibrators and a regression model to calculate absolute quantitative values for unknowns. The reference standard is fully characterized and representative of the biomarker. A relative quantitative assay uses a response–concentration calibration with reference standards that are not fully representative of the biomarker. A quasi-quantitative assay does not employ a calibration standard but has a continuous response that can be expressed in terms of a characteristic of the test sample. Qualitative (categorical) assays can either be described as ordinal (reliant on discrete scoring scales, such as those used in immunohistochemistry) – or nominal (pertaining to a yes/no situation; e.g. the presence or absence of a gene product) [18,19,29]. Although such definitions can never encapsulate every potential biomarker methodology and are by no means accepted in all disciplines, they serve as a guideline for planning performance verification and method validation strategies. Table 1 represents a consensus position on which performance parameters should be investigated for the different classes of biomarker assay, and these are discussed in more detail below under separate headings.

Validation of definitive quantitative biomarker assays

Definitive quantitative methods are less often available in biomarker research but for many represent the ultimate goal for a biomarker assay [30]. Examples include mass spectrometric analysis [31,32] and well-characterized ligand-binding assays (LBAs) [33]. Regardless of the use put to the data, the objective of a definitive quantitative method is to be able to determine as accurately as possible the unknown concentrations of the biomarker in the patient samples under investigation [25]. In this context, analytical accuracy is key and is represented by the total error in the method, consisting of the sum of the systematic error component (bias) and the random error component (intermediate precision) [23,34]. Intermediate precision has to take account of all relevant sources of variation affecting the results (e.g. day, analyst, analytical platform or batch) [10].

TABLE 1

Recommended biomarker assay performance parameters

Performance characteristic	Definitive quantitative	Relative quantitative	Quasi-quantitative	Qualitative
Accuracy	✓			
Trueness (bias)	✓	✓		
Precision	✓	✓	✓	
Reproducibility				✓
Sensitivity	✓ LLOQ	✓ LLOQ	✓	✓
Specificity	✓	✓	✓	✓
Dilution linearity	✓	✓		
Parallelism	✓	✓		
Assay range	✓ LLOQ–ULOQ	✓ LLOQ–ULOQ	✓	
Reagent stability	✓	✓	✓	✓
Sample stability	✓	✓	✓	✓

Internationally recognized performance standards have been established for bioanalytical methods [14,15,35]; however, these were devised primarily by the pharmaceutical industry for small molecule analysis. A study of both precision (% coefficient of variation, or CV) and accuracy (mean % deviation, or bias, from nominal concentration) is required. During the pre-study phase of method validation, precision and accuracy of repeat analyses of the validation samples (VS) are expected to vary by less than 15%, except at the lower limit of quantitation (LLOQ), where 20% is allowable (15/20). When conducting in-study patient sample analysis, quality control samples (QCs) should be employed at three different concentrations in duplicate. The analytical run is accepted as valid when at least 67% (4/6) of the QCs fall within 15% of their nominal values (the 4:6:15 rule) [14,15,36]. Such standards have also been applied to pharmacokinetic studies of macromolecules, where greater leeway is granted within the 4–6–X acceptance rule at either 25% or 30% [34,37–39].

In biomarker method validation, current recommendations also indicate that acceptance criteria for precision and accuracy should be set at a fixed value during pre-study validation [18]. Here, however, more flexibility is allowed: each assay can be evaluated on a case per case basis, with 25% being the default value (30% at LLOQ). Likewise, a similar attitude can be adopted in determining acceptance limits for QCs during patient sample analysis, either in terms of a 4–6–X rule or through adoption of confidence intervals [18,19,29].

Although fixed performance standards are necessary, by their nature they are arbitrary and do not necessarily relate to the intrinsic properties of the assay under investigation or, more importantly, its purpose. The suitability of applying fixed performance criteria in the absence of any statistical evaluation of whether they are relevant to the assay under investigation has been challenged [40,41]. Adopting a 4–6–X rule of acceptability for the QCs means that potentially 33% of the patient samples will also not fall within the acceptance limits. Indeed, because patient samples are usually more heterogeneous matrices than QCs, this value might even be higher. Thus, researchers have questioned whether a method can be considered fit for purpose on the basis of a 4–6–X rule [24]. In a continuing series of original papers, the Societe Francaise des Sciences et Techniques Pharmaceutiques (SFSTP) has developed fit-for-purpose validation of quantitative analytical procedures based on an 'accuracy profile' [22,23,26,42–45]. The accuracy profile takes account of total error (bias and intermediate precision), a pre-set acceptance limit that the user defines (e.g. 20%) and produces a plot of the 'β-expectation tolerance interval' that displays the confidence interval (e.g. 95%, equating to a 5% risk) for future measurements. Effectively, the accuracy profile enables researchers to visually check whether 95% of future values will fall within the chosen acceptance limit of 20%. However, any acceptance limit, confidence interval or level of risk can be represented in the accuracy profile.

To construct an accuracy profile, it is essential that reliable measurements are recorded in the experimental determination of total error. The SFSTP recommend (as a minimum) that 3–5 different concentrations of calibration standards and 3 different concentrations of VS (representing high, medium and low points on the calibration curve) are run in triplicate on 3 separate days, totalling 45 (standard) plus 27 (VS) independent solutions [22,25,26]. Biomarker methods can require a greater number of calibration standards owing to nonlinearity with a concomitant increase in the number of VS. SFSTP also recommended that several different fits to the calibration standards are assessed – because this has a major bearing on accuracy profiles – and that back-calculated values of the calibration standards are used in the calculations of the β-expectation tolerance interval. A full mathematical treatment of the accuracy profile is beyond the scope of this review, but the interested reader is referred to the relevant SFSTP publications.

Other important performance parameters for a definitive quantitative biomarker assay such as sensitivity, dynamic range, LLOQ and upper limit of quantitation (ULOQ) can also be obtained from the accuracy profile. These last two terms are usually defined as the lowest and highest concentrations that can be quantitated with an acceptable level of precision and accuracy (bias). Sample and reagent integrity should also be carefully assessed during method validation for every category of biomarker assay, including studies on specimen stability during collection, storage and analysis; here, it is essential to conduct these studies in authentic patient samples

[16,46,47]. Although specificity, dilution linearity and parallelism should not be overlooked, these parameters are less problematic with a definitive quantitative biomarker because the VS should be identical (or close to identical) in composition to patient samples and should behave in a very similar manner. Specificity, dilution linearity and parallelism will feature in more detail below in the discussion of relative quantitative assays.

Validation of relative quantitative biomarker assays

The LBA is considered by many to be the archetypical quantitative assay for endogenous (protein) macromolecular biomarkers [33] and will act as the focus for discussion in this section. LBAs are available in many different formats, from single analyte sandwich ELISA to diverse multiplex platforms such as Meso-Scale Discovery and Searchlight, and from micro-bead and flow cytometry-based systems such as Luminex beads and Bio-Plex to micro-array ELISA. Specific validation issues associated with multiplex, micro-bead or micro-array systems are addressed in more detail in a series of recent publications [48–51].

To be considered an absolute quantitative method, the reference standard and the matrix must be well defined and representative of the biomarker and the patient sample. Because most biomarker ELISA ligands are endogenous substances, an analyte-free matrix (either to perform specificity studies on or to use as a resource to construct a calibration curve) is usually not available. Access to a fully characterized form of the biomarker to act as a certified calibration standard is also limited [37]. Most available biomarker LBAs, therefore, fall into the category of relative quantitation because they are calibrated with recombinant proteins or peptide standards reconstituted in a surrogate matrix [18].

LBA as a relative quantitative technique is associated with a panoply of specificity issues [40]. Biotransformation (*in vivo* or even *in situ*) precipitated by a variety of factors such as protease or caspase degradation, chemical instability (methionine oxidation, de-amidation or disulfide bond cleavage) or even bacterial contamination can introduce new forms of the biomarker into samples with ill-defined behaviour in the ELISA assay [52,53]. Complexation of the ligand with a soluble receptor, protein aggregation, folding or unfolding of the ligand and insolubility can mask or reveal antibody epitope binding sites [54], which can manifest as a decrease or an increase in concentration [55,56]. Cross-reactivity with closely related protein or peptide moieties is always difficult to characterize fully and eliminate [41]. Finally, abnormalities in blood chemistry (e.g. lipemia) are more likely to adversely affect epitope recognition in cancer patient samples than in QCs made up in 'cleaner' matrices such as plasma or sera from healthy controls [41,57].

LBAs are also highly dependent on the integrity of reagents such as antibodies, which are themselves derived from biologic sources and are subject to their own problems of supply, quality control and stability. Target ranges of the biomarker in the disease group of interest are often unknown and thus expectations are more difficult to define in advance. With a relative quantitative assay, issues of parallelism take on much greater importance. LBAs are also susceptible to non-dilution linearity because antibody- and ligand-binding affinities can vary considerably in different media and the presence of heterophilic antibodies can result in false positive results [41]. To stress the point, several cross-platform studies involving LBA technologies including Multiplex, ELISA, MSD and Luminex have shown that the concentration differences they report with the same samples can be as great as twofold to fivefold [51,58,59].

Resolution of many of the issues cited above is often impossible to achieve purely by studying the performance of the assay during pre-study validation with VS/QCs and requires the analysis of patient samples, an accumulation of clinical data and positive correlation of biomarker concentrations to clinical characteristics.

In the case of a relative quantitative assay, only precision and bias can be evaluated during pre-study validation, not accuracy. Here, it is important to add the caveat that depending on the nature of the calibration standards and matrix of choice, precision and bias determined in VS and QCs might reflect only poorly the true analytical behaviour of the assay in patient samples. Because calibration curves for most LBAs are nonlinear, the AAPS recommends that at least eight to ten different non-zero concentrations should be chosen, with the possibility of a higher density of concentrations at the high and low end to act as anchor points [34,39]. These should be run on three to six separate occasions to establish the most appropriate calibration model. Careful attention should be paid to the curve fitting routine such as 4 or 5-PL and to weighting. The working effective range of the assay should be based on the precision profile where the deviation from the line of best fit to the calibration curve for back-calculated values (% relative error) should lie within an acceptance limit of 10–20% [60]. During pre-study validation, at least five different concentrations of VS – including LLOQ, 3× LLOQ, mid-range, high and ULOQ – should be analyzed in duplicate on at least six different runs. To be considered valid, the LBA should deliver inter-batch precision and bias of $<\pm 20\%$ for each parameter, except at the LLOQ and the ULOQ, where it should be $\pm 25\%$ (20/25); with total error, $<\pm 30\%$ and $\pm 40\%$ (30/40) at the LLOQ and ULOQ [61]. Similar acceptance limits were recommended for in-study validation with QCs, but here only three different concentrations were required to be run in duplicate and a 4–6–X rule used. Although these recommendations are for LBA analysis of macromolecular therapeutic candidates in support of pharmacokinetic and toxicokinetic regulatory studies, they have been largely adopted as integral to performance verification in biomarker method validation but with allowances to extend acceptance criteria if scientifically justified [18,19,29].

Of interest, in a survey that was carried out at the Third AAPS/FDA Bioanalytical Workshop, it was found that 23% of LBA respondents followed a 15/20 rule, 42% a 20/25 rule and 2% a 30/30 rule, whereas 23% used other criteria including statistically based approaches [61]. It is our belief that in biomarker method validation, especially with relative quantitative methods, acceptance criteria should be based on both total error and a confidence interval of 95% [18,19,29,62]. For reasons stated above, the 4–6–X rule should be avoided. We have found with biomarker LBAs that often, in excess of 50–60 repeat measurements of VS/QCs run over weeks are required to reveal the true value of total error in a cumulative plot of intermediate precision (Backen and J.C., unpublished).

In fit-for-purpose biomarker method validation, specificity is defined as the ability to measure the macromolecule in the

presence of other components in the assay matrix. There are two types of non-specificities: specific non-specificity and non-specific non-specificity [60]. Specific non-specificity (or cross-reactivity) can result from interference by macromolecules structurally related to or structurally derived from the biomarker (such as a degradation product, a peptide fragment or a close structural analogue). Non-specific non-specificity (matrix effect) arises from interferences from unrelated species and matrix components (such as heterophilic antibodies) but can often be eliminated by dilution of sample in an appropriate buffer. As explained above, it remains a constant challenge in biomarker research to obtain the relevant test matrices and prove specificity conclusively [10]. According to the AAPS, specificity requires evaluation of concentration–response relationships of both spiked and non-spiked samples obtained from six to ten different sources, preferably patient derived.

Recently, incurred (patient) sample reanalyses as quality controls have been strongly recommended in bioanalysis as a more rigorous test of assay reproducibility [63]. Such an approach has even greater relevance in biomarker analysis, both for increasing confidence in the reproducibility of values and in addressing several non-specificity issues [41]. However, regulatory issues such as ethical approval and patient informed consent would need to be obtained to conduct such an analysis in the UK and the EU in accordance with clinical trial regulations.

The importance of dilution linearity and especially parallelism in the verification of the performance of relative quantitative assays such as LBAs cannot be overemphasized. Dilution linearity is normally studied with spiked QCs during pre-study method validation. Care has to be taken in the choice of matrix to act as the diluent [64]. Parallelism requires access to patient samples and is normally evaluated during in-study validation, provided ethical permission and patient consent are obtained [34]. Although similar to dilution linearity and often confused with this parameter, parallelism is assessed using multiple dilutions of study samples that ideally fall on the quantitative range of the calibration curve – starting at the high end (C_{max}) [60]. Parallelism is dependent on both dilution linearity of patient samples and comparison of the concentration–response relationship of the calibration standards versus the patient samples. Because calibration standards are probably not representative of patient samples, non-parallelism could particularly affect relative quantitative assays.

Parallelism can be conducted with either individual patient samples or a pool of patient samples. Each approach has its pros and cons [19]. There are basically two ways of representing parallelism: the first is as a plot of measured concentrations for the patient samples against 1/dilution factor using log scales. A linear regression analysis is performed using the back-calculated concentration for each dilution of the patient samples, and acceptance criteria can be based either on correlation coefficients or on a statistical acceptance criteria of <20% CV for the deviation of each dilution from the line of best fit [38]. Alternatively, a plot of measured concentrations for the patient samples × dilution against 1/dilution can be constructed. This should yield a flat line so that the CV amongst the recovered concentrations at different dilutions can be used to verify parallelism; here, a CV of <30% is reported as the acceptance limit [60].

Characterization of the stability of biomarkers for analysis by LBA is essential [65]. In the good laboratory practice environment, extensive characterization of sample stability at storage, handling and processing temperature(s) is required by the regulators [46], and these should be conducted in a matrix that mimics the characteristics of the test samples [47]. Use of recombinant proteins in surrogate matrices or even a sample matrix that has been altered is considered less acceptable. Ideally, stability studies should be conducted with incurred (authentic) patient samples [56,64]. To determine significant instability, a change in concentration greater than 2 × total error in the assay (2σ, 95% confidence interval) is required and that two consecutive time points fall out with this limit [66].

Validation of quasi-quantitative biomarker assays

Quasi-quantitative assays lack calibration but report numerical values (e.g. detector response) as a characteristic of the sample. Such techniques include quantitative RT-PCR (qRT-PCR) or a poorly defined ELISA. Precision, specificity, sensitivity and the dynamic range of the assay form the core of the performance parameters that should be verified during pre-study validation.

As an example, we have conducted a fit-for-purpose validation of qRT-PCR using an amplification refractory mutation system assay and Scorpion probes for mutation detection in K-RAS, PI3K and EGF-R in free circulating DNA [67–69]. Because the ultimate purpose of this assay is to detect mutations in the serum or plasma of cancer patients, validation should not be considered complete until it is demonstrated that the assay does indeed detect the mutation in cancer patients with acceptable clinical sensitivity (detects the mutation when present) and clinical specificity (does not detect the mutation when absent) [67,68].

Several different positive and negative controls were incorporated into the assay (a key performance characteristic of both quasi and qualitative assays) as a quality assurance measure, including PCR control, an external QC and the in-kit mixed standard control. Where possible, pools of patient-derived samples should also be used to generate positive and negative controls, if it can be confirmed independently that these subjects either contain or lack the mutation under investigation. The CLIA regulations in the USA require only verification (i.e. confirmation) of performance for (FDA) approved molecular clinical tests (such as qRT-PCR) if the test system has been previously validated by the manufacturer and is used without any modifications [12]. Our assessment on the precision of these in vitro diagnostic compliant tests (K-RAS and EGF-R) was also conducted on a confirmatory basis, requiring only three to five assays to fall within the manufacturer's specification. The acceptance limit was set at 2 × the change in threshold value (ΔCT) between mutated and non-mutated DNA. All precision data obtained during verification, validation and patient sample analysis were then incorporated into a QC data base for ongoing quality control and competency testing, in keeping with the QC monitoring of assays in routine clinical use [22,26]. The QC data were especially important in the case of these quasi-quantitative assays to mitigate against the lack of absolute numbers to monitor assay performance against.

In another example of fit-for-purpose validation, a quasi-quantitative ELISA measuring DNA nucleosomes (nDNA) was evaluated as a biomarker of cell death (Cell Death Detection ELISA^{Plus} from

ROCHE Diagnostics, Ltd., Burgess Hill, UK). The sandwich ELISA is supplied only with a positive control and is not calibrated against a standard, so assay readout is in absorbance units generated by the microplate reader. Our fit-for-purpose approach focused on demonstrating utility in the analysis of clinical trial samples. Although quality controls were prepared by titration of the in-kit positive control, because these were reconstituted in buffer their value was limited to acting as an aid to performance verification. Extensive stability studies were performed on DNA nucleosomes made up in buffer to replicate the QCs or spiked into serum or plasma to replicate patient samples. Sample collection proved crucial to the analysis of nDNA. Careful handling of whole blood was required to avoid haemolysis and artifactual production of nDNA, and to preserve stability centrifugation to isolate serum or plasma was necessary as soon as practicable before storage at −80°C. For the in-study analysis of patient samples, we recommend including four to six replicates of the positive control QC to verify the assay is working correctly, without a strict limit placed on acceptance criteria. More importantly, we believe it is essential to include up to six patient samples for incurred sample reanalysis, as the primary test of quality control and reproducibility. Here, a fixed acceptance limit is adopted of $<\pm 30\%$ compared to the values determined in the previous assay, in the absence of sufficient data to apply confidence intervals.

Our recent application of this assay to the analysis of clinical specimens has confirmed clinical utility both as a pharmacodynamic marker [70] and as a predictive biomarker of response to cancer chemotherapy [71], confirming several previous publications [72–75]. Nonetheless, there are limitations with a quasi-quantitative assay: batch-to-batch QC and competency testing are difficult to evaluate, and imprecision is usually greater than with a typical relative quantitative ELISA. Paradoxically, however, biomarker assays such as quasi-quantitative or qualitative techniques that require verification of fewer performance characteristics (Table 1) have perhaps an even greater requirement to be fast tracked into the clinic to complete the process of validation.

Validation of qualitative biomarker assays

Qualitative biomarker assays such as immunohistochemistry or fluorescence *in situ* hybridization are often positioned at the diagnostic end of the biomarker spectrum. This combination should dictate that verification of performance only contributes a small fraction of the total experimentation required to constitute a validated assay, with clinical investigations accounting for the vast majority of the data. Use of terms such as 'precision' and 'accuracy' are not considered appropriate with qualitative assays [12]. Positive and negative controls are the mainstay to confirm assay performance, whereas to increase reliability, more than one trained investigator should score images [18,19].

Proving analytical specificity (that the assay does not detect the biomarker when absent) can require access to resources such as knock-out mice, expressed cell lines or clinical specimens that are not readily available. We recommend a risk-based approach to specificity, conducting more limited studies with resources that are readily available to at least reduce the level of uncertainty. Sensitivity is usually set at the expression level at which informative positive results are obtained 95% of the time, but again, to conduct a thorough investigation of this parameter requires access to the appropriate resources [12]. Assessment of reproducibility involves repeat analysis of multiple patient samples to demonstrate consistency. The Clinical Laboratory Standards Institute evaluation protocol for a qualitative test recommends analysis of a minimum of 50 positive and 50 negative specimens run over 10–20 days [76]. If the test kit or a key reagent such as an antibody has already undergone extensive validation by the vendor, however, then confirmation of performance within manufacturer's specifications by the laboratory of interest should be adequate to satisfy the regulators. It is recommended that the validation data produced by the manufacturer is obtained by the laboratory for scrutiny.

The different phases of fit-for-purpose biomarker assay validation

Biomarker method validation can be viewed to occur in discrete stages, each with a specific purpose, defined goal and end product [15,18,19] (Table 2). Perhaps the most important stage is the first, from which all others inevitably flow, where definition of global purpose and judicious selection of the candidate assay occurs. During method development (stage 2) the goal is to assemble all the appropriate reagents and components, and it is only at this stage that the final classification of the assay into one of the five categories will occur. The method validation plan is also constructed at this stage. Stage 3 is the pre-study experimental phase of performance verification culminating with the all-important evaluation of fitness for purpose. The end product of this stage is an analytical report and standard operating procedure to take forward to in-study patient sample analysis. In-study validation

TABLE 2

Stages in biomarker assay validation

Stage	Description	Main purpose	Evaluation	End product
1	Define use, seek appropriate assay	Characterize the clinical and experimental aims and objectives	Is there an assay potentially fit for the purpose?	Candidate assay
2	Method development	Assemble all components; develop method and perform preliminary validation	Go–no-go decision	Validation plan
3	Pre-study validation	Run validation samples; characterize performance	Performance verification	A validated method, report and standard operating procedure
4	In-study validation	QC monitoring; identify patient sampling issues	Fit-for-purpose	Valid patient data
5	Routine use	QC monitoring; proficiency testing; batch-to-batch QC	Continuous improvement	Continued use

(stage 4) enables further assessment of fitness for purpose, collection of QC data and identification of patient sampling issues. Only then can stage 5 be reached, in which the assay enters into routine use. Here, QC monitoring continues and proficiency testing and batch-to-batch quality control issues can be fully explored. The overarching philosophy of the whole process is one of continual improvement, which might precipitate a series of iterations that can lead back to redefinition of purpose (stage 1), modification of experimental procedures (stage 2), further characterization of assay performance (stage 3) or even re-assessment of patient sampling issues (stage 4) [19,29].

Validation of commercially available biomarker assays

It has been acknowledged in this review that commercially available assays generally used for the purpose of a diagnostic test are treated distinctly during method validation, provided that the test is approved by a regulatory agency and is not subject to change. By their very nature, most commercial biomarker assays are experimental tools and not approved and, therefore, fitness of purpose has to be established for an alternate use in biomarker research, especially involving the analysis of patient samples. In that scenario (Fig. 3), we have developed a generic validation strategy for commercially available LBAs [50]. The approach, which is essentially a confirmation that the assay performs within specifications (either the manufacturer's or set in-house), uses QCs and consists of two stages. In the first (Fig. 3), the imprecision in the QCs (as a % CV) is determined experimentally by analysing four replicates of up to three different QC concentrations in five different assays, preferably run on separate days. These data are then used to set target, but nevertheless preliminary, acceptance limits at the 95% confidence interval (2σ or 2× the CV% for imprecision) against which the performance of future assays is evaluated. To be considered acceptable to take further forward to analysis of clinical trial samples, the analyst must demonstrate that three additional assays fall within the target CVs for the QCs. Once stage 1 is complete, an interim study report is written and the validation progresses to patient sample analysis and stage 2, where all the key components of fit-for-purpose validation are studied (such as sample stability and handling issues, dilution linearity and parallelism and definition of biomarker target concentrations). QC monitoring continues until the plateau phase in the cumulative plot of precision is reached (representing the total true error in the measurement of the CVs). Acceptance criteria are then fixed but still at the 95% confidence interval and only changed if batch-to-batch issues arise. The success of this validation approach relies heavily on an accurate determination of the total error associated with measurement of the QCs.

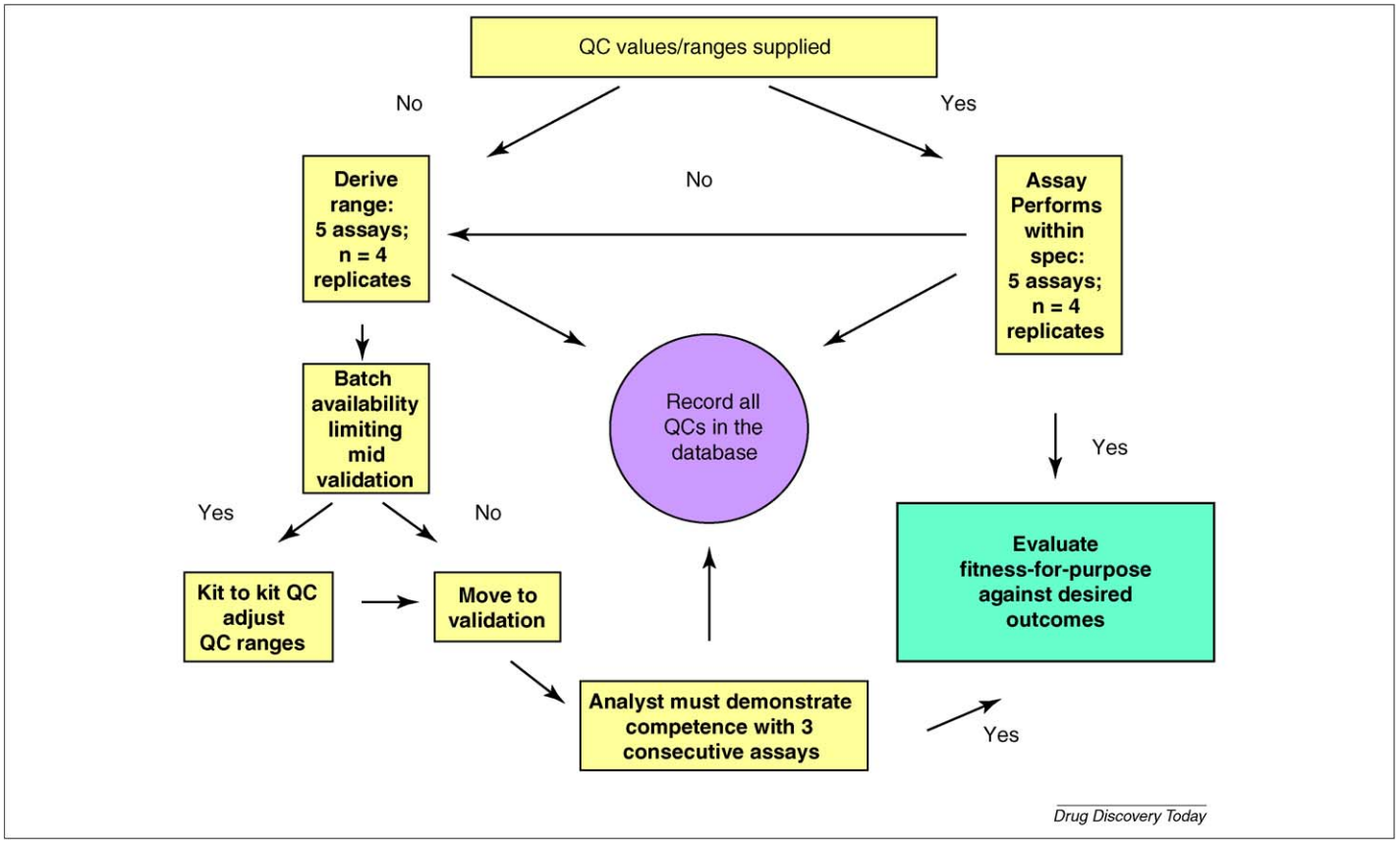

FIGURE 3

Fit-for-purpose validation of a generic commercially available ligand-binding assay. In the first stage of the fit-for-purpose approach, the goal is essentially to confirm that the assay performs within specifications (either manufacturer's or set in-house) and uses quality control (QC) samples. The imprecision in the QCs (as a % CV) is determined experimentally by analysing four replicates of up to three different QC concentrations in five different assays, preferably run on separate days. These data are then used to set target acceptance limits at the 95% confidence interval, against which the performance of future assays is evaluated. To be considered acceptable to take further forward to analysis of clinical trial samples, the analyst must demonstrate that three additional assays fall within the target CVs for the QCs. For the second stage, see text.

In conclusion, fit-for-purpose validation provides researchers with a sensible approach to biomarker method validation that tailors the burden of proof to take account of both the nature of technology used and the impact of the result on the future development or use of the drug.

Acknowledgements

Cancer Research UK and Experimental Cancer Medicine Centre are acknowledged for funding the present work. Alison Backen, Grace Hampson, Nigel Smith and Fouziah Butt are thanked for their technical input and for many useful discussions.

References

1 Zerhouni, E. (2003) Medicine. The NIH roadmap. *Science* 302, 63–72
2 McShane, L.M. *et al.* (2009) Effective incorporation of biomarkers into phase II trials. *Clin. Cancer Res.* 15, 1898–1905
3 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–715
4 Carden, C.P. *et al.* (2009) From darkness to light with biomarkers in early clinical trials of cancer drugs. *Clin. Pharmacol. Ther.* 85, 131–133
5 Wagner, J.A. *et al.* (2007) Biomarkers and surrogate end points for fit-for-purpose development and regulatory evaluation of new drugs. *Clin. Pharmacol. Ther.* 81, 104–107
6 Carden, C.P. *et al.* (2010) Can molecular biomarker-based patient selection in Phase I trials accelerate anticancer drug development? *Drug Discov. Today* 15, 88–97
7 Pepe, M.S. *et al.* (2001) Phases of biomarker development for early detection of cancer. *J. Natl. Cancer Inst.* 93, 1054–1061
8 Maruvada, P. and Srivastava, S. (2006) Joint National Cancer Institute–Food and Drug Administration workshop on research strategies, study designs, and statistical approaches to biomarker validation for cancer diagnosis and detection. *Cancer Epidemiol. Biomarkers Prev.* 15, 1078–1082
9 Jain, R.K. *et al.* (2009) Biomarkers of response and resistance to antiangiogenic therapy. *Nat. Rev. Clin. Oncol.* 6, 327–338
10 Cummings, J. *et al.* (2008) Biomarker method validation in anticancer drug development. *Br. J. Pharmacol.* 153, 646–656
11 FDA, (2005) *Guidance for Industry Pharmacogenomic Data Submissions*. http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm126957.pdf
12 Jennings, L. *et al.* (2009) Recommended principles and practices for validating clinical molecular pathology tests. *Arch. Pathol. Lab. Med.* 133, 743–755
13 Chau, C.H. *et al.* (2008) Validation of analytic methods for biomarkers used in drug development. *Clin. Cancer Res.* 14, 5967–5976
14 Shah, V.P. *et al.* (1991) Analytical methods validation: bioavailability, bioequivalence and pharmacokinetic studies. Conference report. *Eur. J. Drug Metab. Pharmacokinet.* 16, 249–255
15 Shah, V.P. *et al.* (2000) Bioanalytical method validation – a revisit with a decade of progress. *Pharm. Res.* 17, 1551–1557
16 FDA, (2001) *Guidance for Industry: Bioanalytical Method Validation.* http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070107.pdf
17 Shah, V.P. (2007) The history of bioanalytical method validation and regulation: evolution of a guidance document on bioanalytical methods validation. *AAPS J.* 9, E43–E47
18 Lee, J.W. *et al.* (2005) Method validation and measurement of biomarkers in nonclinical and clinical samples in drug development: a conference report. *Pharm. Res.* 22, 499–511
19 Lee, J.W. *et al.* (2006) Fit-for-purpose method development and validation for successful biomarker measurement. *Pharm. Res.* 23, 312–328
20 17025, I.I., (2005) *General Requirements for the Competence of Testing and Calibration Laboratories*. ISO http://www.iso.org/iso/catalogue_detail.htm?csnumber=39883
21 9000, I., (2005) *Quality Management Systems – Fundamentals and Vocabulary*. International Organization for Standardization http://www.iso.org/iso/catalogue_detail.htm?csnumber=42180
22 Feinberg, M. *et al.* (2004) New advances in method validation and measurement uncertainty aimed at improving the quality of chemical data. *Anal. Bioanal. Chem.* 380, 502–514
23 Hubert, P. *et al.* (2004) Harmonization of strategies for the validation of quantitative analytical procedures. A SFSTP proposal – Part I. *J. Pharm. Biomed. Anal.* 36, 579–586
24 Boulanger, B. *et al.* (2009) A risk-based analysis of the AAPS conference report on quantitative bioanalytical methods validation and implementation. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* 877, 2235–2243
25 Rozet, E. *et al.* (2007) Using tolerance intervals in pre-study validation of analytical methods to predict in-study results. The fit-for-future-purpose concept. *J. Chromatogr. A* 1158, 126–137
26 Feinberg, M. (2007) Validation of analytical methods based on accuracy profiles. *J. Chromatogr. A* 1158, 174–183
27 Kelloff, G.J. and Sigman, C.C. (2005) New science-based endpoints to accelerate oncology drug development. *Eur. J. Cancer* 41, 491–501
28 Ludwig, J.A. and Weinstein, J.N. (2005) Biomarkers in cancer staging, prognosis and treatment selection. *Nat. Rev. Cancer* 5, 845–856
29 Lee, J.W. *et al.* (2007) Biomarker assay translation from discovery to clinical studies in cancer drug development: quantification of emerging protein biomarkers. *Adv. Cancer Res.* 96, 269–298
30 Cummings, J. *et al.* (2009) Quantitative analysis of biomarkers by LC–MS/MS. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* 877, 1221
31 Barnidge, D.R. *et al.* (2004) Absolute quantification of the model biomarker prostate-specific antigen in serum by LC–MS/MS using protein cleavage and isotope dilution mass spectrometry. *J. Proteome Res.* 3, 644–652
32 Fenselau, C. (2007) A review of quantitative methods for proteomic studies. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* 855, 14–20
33 van der Merwe, D.E. *et al.* (2007) Mass spectrometry: uncovering the cancer proteome for diagnostics. *Adv. Cancer Res.* 96, 23–50
34 DeSilva, B. *et al.* (2003) Recommendations for the bioanalytical method validation of ligand-binding assays to support pharmacokinetic assessments of macromolecules. *Pharm. Res.* 20, 1885–1900
35 Peters, F.T. *et al.* (2007) Validation of new methods. *Forensic Sci. Int.* 165, 216–224
36 Bansal, S. and DeStefano, A. (2007) Key elements of bioanalytical method validation for small molecules. *AAPS J.* 9, E109–E114
37 Miller, K.J. *et al.* (2001) Workshop on bioanalytical methods validation for macromolecules: summary report. *Pharm. Res.* 18, 1373–1383
38 Findlay, J.W. *et al.* (2000) Validation of immunoassays for bioanalysis: a pharmaceutical industry perspective. *J. Pharm. Biomed. Anal.* 21, 1249–1273
39 Smolec, J. *et al.* (2005) Bioanalytical method validation for macromolecules in support of pharmacokinetic studies. *Pharm. Res.* 22, 1425–1431
40 Findlay, J.W. (2008) Specificity and accuracy data for ligand-binding assays for macromolecules should be interpreted with caution. *AAPS J.* 10, 433–434
41 Findlay, J.W. (2009) Some important considerations for validation of ligand-binding assays. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* 877, 2191–2197
42 Boulanger, B. *et al.* (2003) An analysis of the SFSTP guide on validation of chromatographic bioanalytical methods: progress and limitations. *J. Pharm. Biomed. Anal.* 32, 753–765
43 Hubert, P. *et al.* (2007) Harmonization of strategies for the validation of quantitative analytical procedures. A SFSTP proposal – part II. *J. Pharm. Biomed. Anal.* 45, 70–81
44 Hubert, P. *et al.* (2007) Harmonization of strategies for the validation of quantitative analytical procedures. A SFSTP proposal – part III. *J. Pharm. Biomed. Anal.* 45, 82–96
45 Hubert, P. *et al.* (2008) Harmonization of strategies for the validation of quantitative analytical procedures: a SFSTP proposal part IV. Examples of application. *J. Pharm. Biomed. Anal.* 48, 760–771
46 James, C.A. and Hill, H.M. (2007) Procedural elements involved in maintaining bioanalytical data integrity for good laboratory practices and regulated clinical studies. *AAPS J.* 9, E123–E127
47 Nowatzke, W. and Wood, E. (2007) Best practices during bioanalytical method validation for the characterisation of assay reagents and the evaluation of analyte stability in assay standards, quality controls and study samples. *AAPS J.* 9, E117–E122
48 Gonzalez, R.M. *et al.* (2008) Development and validation of sandwich ELISA microarrays with minimal assay interference. *J. Proteome Res.* 7, 2406–2414
49 Ling, M.M. *et al.* (2007) Multiplexing molecular diagnostics and immunoassays using emerging microarray technologies. *Expert Rev. Mol. Diagn.* 7, 87–98
50 Backen, A.C. *et al.* (2009) 'Fit-for-purpose' validation of SearchLight multiplex ELISAs of angiogenesis for clinical trial use. *J. Immunol. Methods* 342, 106–114
51 Chowdhury, F. *et al.* (2009) Validation and comparison of two multiplex technologies. Luminex and Mesoscale Discovery, for human cytokine profiling. *J. Immunol. Methods* 340, 55–64
52 Maity, H. *et al.* (2009) Effects of pH and arginine on the solubility and stability of a therapeutic protein (Fibroblast Growth Factor 20): relationship between solubility and stability. *Curr. Pharm. Biotechnol.* 10, 609–625

53 Mahler, H.C. *et al.* (2009) Protein aggregation: pathways, induction factors and analysis. *J. Pharm. Sci.* 98, 2909–2934
54 Wu, F.T. *et al.* (2009) A compartment model of VEGF distribution in humans in the presence of soluble VEGF receptor-1 acting as a ligand trap. *PLoS ONE* 4, e5108
55 Nayeri, F. *et al.* (2002) Sample handling and stability of hepatocyte growth factor in blood samples. *Cytokine* 19, 201–205
56 Cummings, J. *et al.* (2007) Qualification of M30 and M65 ELISAs as surrogate biomarkers of cell death: long term antigen stability in cancer patient plasma. *Cancer Chemother. Pharmacol.* 60, 921–924
57 Deligezer, U. *et al.* (2006) Circulating fragmented nucleosomal DNA and caspase-3 mRNA in patients with lymphoma and myeloma. *Exp. Mol. Pathol.* 80, 72–76
58 Urbanowska, T. *et al.* (2006) Protein microarray platform for the multiplex analysis of biomarkers in human sera. *J. Immunol. Methods* 316, 1–7
59 Toedter, G. *et al.* (2008) Simultaneous detection of eight analytes in human serum by two commercially available platforms for multiplex cytokine analysis. *Clin. Vaccine Immunol.* 15, 42–48
60 Kelley, M. and DeSilva, B. (2007) Key elements of bioanalytical method validation for macromolecules. *AAPS J.* 9, E156–E163
61 Viswanathan, C.T. *et al.* (2007) Quantitative bioanalytical methods validation and implementation: best practices for chromatographic and ligand binding assays. *Pharm. Res.* 9, E30–E42
62 Westgard, J.O. (1994) Selecting appropriate quality-control rules. *Clin. Chem.* 40, 499–501
63 Fast, D.M. *et al.* (2009) Workshop report and follow-up – AAPS Workshop on current topics in GLP bioanalysis: assay reproducibility for incurred samples – implications of Crystal City recommendations. *AAPS J.* 11, 238–241
64 Greystoke, A. *et al.* (2008) Optimisation of circulating biomarkers of cell death for routine clinical use. *Ann. Oncol.* 19, 990–995
65 Aziz, N. *et al.* (1999) Variables that affect assays for plasma cytokines and soluble activation markers. *Clin. Diagn. Lab. Immunol.* 6, 89–95
66 Hoffman, D. *et al.* (2009) Statistical methods for assessing long-term analyte stability in biological matrices. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* 877, 2262–2269
67 Board, R.E. *et al.* (2010) Detection of PIK3CA mutations in circulating free DNA in patients with breast cancer. *Breast Cancer Res. Treat.* 120, 461–467
68 Board, R.E. *et al.* (2008) Multiplexed assays for detection of mutations in PIK3CA. *Clin. Chem.* 54, 757–760
69 Hodgson, D.R. *et al.* (2010) Circulating tumour-derived predictive biomarkers in oncology. *Drug Discov. Today* 15, 98–101
70 Dean, E. *et al.* (2009) Phase I trial of AEG35156 administered as a 7-day and 3-day continuous intravenous infusion in patients with advanced refractory cancer. *J. Clin. Oncol.* 27, 1660–1666
71 Hou, J.M. *et al.* (2009) Evaluation of circulating tumor cells and serological cell death biomarkers in small cell lung cancer patients undergoing chemotherapy. *Am. J. Pathol.* 175, 808–816
72 Holdenrieder, S. *et al.* (2001) Nucleosomes in serum of patients with benign and malignant diseases. *Int. J. Cancer* 95, 114–120
73 Holdenrieder, S. *et al.* (2001) Nucleosomes in serum as a marker for cell death. *Clin. Chem. Lab. Med.* 39, 596–605
74 Holdenrieder, S. *et al.* (2006) Early and specific prediction of the therapeutic efficacy in non-small cell lung cancer patients by nucleosomal DNA and cytokeratin-19 fragments. *Ann. N. Y. Acad. Sci.* 1075, 244–257
75 Holdenrieder, S. *et al.* (2004) Circulating nucleosomes predict the response to chemotherapy in patients with advanced non-small cell lung cancer. *Clin. Cancer Res.* 10, 5981–5987
76 Wiktor, A.E. *et al.* (2006) Preclinical validation of fluorescence *in situ* hybridization assays for clinical practice. *Genet. Med.* 8, 16–23