Addressing the challenges of biomarker calibration standards in ligand-binding assays: a European Bioanalysis Forum perspective

The analysis of biomarkers by ligand-binding assays offers significant challenges compared with the bioanalysis of small and large molecule drugs. The presence of endogenous analyte is a commonly cited issue. Also the sourcing and application of appropriate calibration or reference standards can present many issues. One of the main challenges is ensuring the continuity and validity of biomarker data when the source or lot number of calibration standard changes within or between studies. Several strategies exist in attempting to deal with this and standardize the biomarker data through the assay life or looking for ways to compare and normalize biomarker data. In this manuscript, the European Bioanalysis Forum view on dealing with calibration standards in biomarker assays is described.

First draft submitted: 27 June 2017; Accepted for publication: 11 August 2017;

Published online: 21 October 2017

Keywords: biomarker • calibration and reference standard • ligand-binding assay

In an era where multifactorial biomarker models continue to increase in complexity, standardization of the analytical measurement process is critical to promote drug product development. Ligand-binding assays (LBA) or immunoassays are used for the majority of biomarker measurements performed during the various stages of drug development.

Due to the analytical principle of LBAs using selective antibody–antigen affinity interactions for detecting the analyte of interest and the external calibration, LBAs do not provide a direct measurement in absolute amounts of analyte (e.g., gravimetry using a balance). LBAs estimate relative analyte concentrations in unknown samples by comparing the sample instrument signal to that from similar treated standard samples with a known target concentration. The prerequisite of such an external calibration of assays is an adequate standard that is identical to the analyte in the test samples, which behaves in the same manner in the

assay procedure and is available in the same well characterized quality for the whole duration of assay use. These criteria are rarely met in the field of protein biomarkers measured by LBAs [1] unlike pharmacokinetic (PK) assays where the drug substance itself is used as the calibration standard.

A challenge with a major impact on the use of quantitative LBAs in drug development is the variability of the calibration standards between manufacturers, between lots from the same manufacturer and even between vials of the identical lot [2]. Commercial suppliers may use reference or calibration standards that are derived from different cell lines and these cell lines may differ in their level of characterization and host cell protein contamination. Furthermore suppliers often assign a numerical value to their calibration standard according to activity results from their respective bioassay. Different suppliers can have LBAs for the same biomarker but they may differ in the selectivity and affinity of the capture

Ulrich Kunz¹, Joanne Goodman², Ulf Loevgren³, Timo Piironen⁴, Karen Elsby⁵, Paul Robinson⁶, Susanne Pihl⁷, Amanda Versteilen⁸, Arjen Companjen⁸, Marianne Scheel Fjording⁹ & Philip Timmerman^{*,10}

¹Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Strasse 65, 88397 Biberach an der Riss, Germany

²MedImmune, Aaron Klug Building, Granta Park, Cambridge, CB21 6GH, UK

³Ferring Pharmaceuticals A/S, Kay Fiskers Plads 11, DK-2300 Copenhagen S, Denmark

⁴Syrinx Bioanalytics Oy, Pansiontie 47, Biohouse D5, FI-20210 Turku, Finland

⁵AstraZeneca R&D Alderley Park, Macclesfield, Cheshire, UK

⁶LGC Ltd, CB7 5WW, UK

⁷Ascendis Pharma A/S, Tuborg Boulevard 5, 2900 Hellerup, Denmark

⁸Janssen Vaccines & Prevention B.V., Fakkellgras 10, 2224JT Katwijk, The Netherlands

⁹Novo Nordisk A/S, Novo Nordisk Park, DK-2760 Måløv, Denmark

¹⁰European Bioanalysis Forum, Belgium

*Author for correspondence: philip@e-b-f.eu

newlands
press part of

fsg

and detection antibodies, leading to huge differences in the concentrations measured by the assay. The production and quality control processes for research grade proteins or kits are not standardized or well controlled as they are for the GMP production of drugs or certified diagnostics; a lot-to-lot deviation of 76% has been observed in a European Bioanalysis Forum (EBF) case study. Another source of result shifts between runs is the vial to vial variability of lyophilized materials after reconstitution. A variability of up to 10% within a lot often is considered acceptable by the manufacturer which can pose issues when combined with inherent assay variance in LBA.

All these sources of variability affect the comparability of sample results measured on different occasions in different runs. The ideal case would be to measure a single study with the same lot of calibration standard only. But what about follow-up studies or long-term studies that last over several years, for example, Phase III or oncology trials? The challenge of a lot-to-lot variability and long-term comparability of biomarkers results is well known in clinical diagnostics and also in biopharmaceutical drug manufacturing. A long-term comparability approach has to be flexible and should cover the option of multiple changes of lots with respect to limited stability information available for standard and reagents, expiry dates of reagents and kits. In order to avoid a drift in results careful monitoring of assay performance and the use of comparator samples becomes important. Possible comparator samples could take the form of an International Standard, if available, an in-house reference standard preparation, banked sample controls or representative data from a suitable population. The latter approach assumes that the intersubject and the longitudinal variability will not vary substantially [3].

Due to the broad range of intended uses of biomarker measurements it is useful to consider a tiered approach, not only for assay validation but also regarding the effort that is required for the comparability of results [4].

There is a lack of detailed discussion in white papers on biomarker assay validation for drug development [5] and the associated challenges. In this paper, the EBF reports back from their internal discussions on the best approaches regarding bioanalysis of biomarkers in support of drug development performed in the regulated bioanalytical environment. These discussions were an integral part of an EBF subteam assigned to provide a perspective on the various approaches and practical implementation of the challenges of lot-to-lot variability of standards and long-term comparability of biomarker results

measured with LBAs. Some important key terms are explained in 'Box 1'.

Perspectives & best practices

Procurement & stability of calibration standards

Surrogate versus endogenous analyte

Biomarkers and their respective standards, come in many forms and from different origins. Many endogenous protein biomarkers appear in various natural isoforms with various post-translational modifications and certain forms that can increase or decrease with different physiological states, nutrition, sex or age. Conversely, biomarker calibration standards are purified, artificially produced compounds, for example low molecular weight molecules, peptides or recombinant proteins from various cell lines. The natural variation of the endogenous analytes and analyte composition could neither be mimicked by a single calibration standard formulation nor a recombinant surrogate standard [7].

The unavailability of standard that is identical to the endogenous analyte leads to the definition of relative quantitative assays in contrast to definitive quantitative assays as defined by Lee *et al.* [8].

Different formulations of calibration standards (solid, lyophilized & stock solution)

Standards can be obtained or produced in different formulations, for example, as solid powder, stock solution or lyophilized. Solid formulations are chosen for chemically synthesized standards and are stable for years if stored in cool, dry conditions and protected from light. This also applies for lyophilized protein standards. Lyophilization from solution together with several auxiliary reagents is often performed for the bulk production of proteins. A reproducible lyophilization process is a challenge and sometimes the reason for significant lot-to-lot as well as vial-to-vial variability.

Standards in stock solutions are generally stored deep frozen in aliquots for ease of use. The storage solution must be formulated in such a way that it promotes stability of the standard even during freezing and thawing. Aliquots of a reference stock solution should be thawed only once and the remaining aliquot volume should be discarded. Aliquots of calibration standard stock solutions have the potential to be stored cooled after thawing if stability investigations are performed [9]. The volume of the aliquots should be sufficient for the daily use but not too small to prevent freeze drying during long-term storage. Size and type of the cryovial also influences freeze drying and small vials with a tight cap are often the best choice.

Box 1. Key terms

The following definitions follow international accepted rules [1].

- **Calibration:** the process of assigning values to unknown samples using a standard
- **International standard:** international reference preparation, WHO International Laboratory for Biological Standards, often prepared by the National Institute for Biological Standards and Control (UK); activity in international units. Can be used for cross-validation of different methods and reference standards for the same biomarker
- **Reference standard:** a validated source of the analyte, which is acknowledged as having appropriate qualities within a specified context with an accepted value. The reference standard may be used as a primary standard for the determination of the effect, activity and/or concentration of the biomarker; a reference standard is known to be available in reproducible quality over long-term or can be banked for long-term. It can be used to bridge various lots of calibration standards and to guarantee comparability of results of a particular assay over long-term
- **Calibration standard:** standard material with nominal content of (surrogate) analyte used to prepare calibration samples and to calibrate the assay run by means of a calibration curve from which the (relative) concentration of the desired analyte in study samples are determined. Use of the calibration standard is under the control of the bioanalytical lab. The calibration standard is also known as the working standard. It should guarantee comparability of results between different runs of a particular assay
- **Calibration samples/calibrators:** a series of test samples prepared from the calibration standard to calibrate the assay run by means of fitting a calibration curve to the obtained response values
- **Reference material:** more general term for all kinds of analytical methods that need to compare determined results of unknown samples to known samples or materials
- **Commercial supply:** reagents or assay kits that can be bought from catalogue, in contrast to reagents that are produced specially on demand of a sponsor or self produced in-house for a particular assay
- **Validation:** process of testing an analytical method whether the method performance is acceptable for the intended use. Validation is a terminology often ambiguously used and verification thereof is required [6]
- **Normalization:** measuring of a test material versus a reference material and adjusting the nominal value of the test material to compensate for different reactivity in the particular assay
- **Trending:** monitoring assay parameters and reactivity of monitoring samples over time to indicate long-term trends and jumps in assay performance maybe due to instability of reagents or lot-to-lot changes

Commercial supplier

International reference standards should ideally be obtained from internationally recognized and specialized institutes, such as the European Directorate for the Quality of Medicine & Healthcare, The National Institute for Biological Standards and Control, US Pharmacopeia, the Expert Committee on Biological Standardization of WHO and the American Type Culture Collection [10].

They are generally intended for standardization of the in-house laboratory reference standards and are routinely used in clinical diagnostics. They are supplied either as highly purified, endogenous protein extracts from blood without contamination by plasma proteins or as a well characterized purified, recombinant proteins. The content is assigned in mass per volume or international units (IU), the latter being an arbitrary unit as the exact content is hard or nearly impossible to determine. They at least provide a common, unitized source of analyte for all the methods available. However, it is common that over time, different lots of International Standards will differ in concentration or IU due to a different source of donor blood.

Material for reference or calibration standards that can be obtained from commercial suppliers, are part of commercial assay kits or have to be produced in-house. Selection of a suitable standard is an important part of the method development or feasibility process. It is recommended to test different production lots, if available, [11] and to request details for the standard from the supplier if the general product description is not sufficient. Replacing the standard of a commercial kit by a separately ordered bulk quantity in order to use it through several lots of kits is one option to reduce lot-to-lot variability. If a single lot of calibration standard is intended to be used, ensure sufficient stock of the same lot number is secured for the required duration.

Reliability of vendor

When sourcing calibration standards (or assay kits) commercially, consideration should be paid to the supplier reliability. Consider their ability to supply the kit for the duration of the study, and the frequency of kit changes made historically. Obtain information relating to reagent characterization, standardization and stability assessments conducted by the vendor, in addition to documentation relating to their quality control pro-

cesses and acceptance criteria. Suppliers will not always notify users of batch production changes; however, such process alterations might have a significant impact on the assay results.

Expiry date

Standards sourced from a commercial supplier, would typically be supplied with an assigned expiry date. Where an expiry date is provided, it is assumed that the stability of the standard is assured until this time. However, in some cases, an expiry date is not assigned or the assigned expiry is based on limited data.

Although it is desirable to use a single lot of calibration standard throughout a study within an assigned expiry date, this is not always possible. In these cases, as an alternative to using multiple lots, it may be justifiable to assign a new expiry or retest date to the standards, or to simply use the calibration standards past the assigned expiry date.

In a survey distributed to the EBF community by members of this topic team, it was found that 35% of all responders stated that they would use a calibration standard past the expiry date of the vendor. In-house setting of expiry dates, that is through stress stability tests or physiochemical characterization is a feasible option for using standards past the assigned expiry although this approach is not commonly used according to the survey (none of the 35% currently take this approach) due to the tremendous effort and the huge demand on material necessary for such GMP-like stability investigations.

If calibration standards are subsequently used past their expiry date, it is recommended to perform as a minimum a trend analysis in order to provide early indication of instability. This early indication may also provide sufficient time to perform comparison or bridging of the standard to a new lot if required. In general, the storage of proteins deep frozen at -70°C , at -150°C or even at both temperatures in parallel is considered suitable for long-term storage and routinely performed for storage of primary reference standards and calibration standards for quality control of biopharmaceuticals and diagnostics. A nonparallel trend in signals from aliquots stored at both temperatures is a sensitive indicator for instability.

Certificate of Analysis

For calibration standards it is best practice for a Certificate of Analysis (CofA) to be obtained and stored with the raw data. The minimum required information on such a CofA would be a unique terminology of the calibration standard, the nominal concentration, the manufacturer and a lot number. Although an expiry date is most often requested by quality assurance units

it is not a prerequisite from a scientific perspective as often the expiry date is not based on stability data and not predictive for the particular vial. The use of a retest date would be an alternative. Additionally, details about the source and identity of the calibration standard, for example cell line, amino acid sequence, buffer or auxiliary reagents might be useful in comparing material from different vendors.

Tools to determine & overcome differences in lots of calibration standards

Normalization

A common strategy to make results comparable and to remove bias is to employ normalization of observed values versus an accepted reference value. In the case of different standard lots this can be implemented by various concepts, for example, when comparing QC samples or calibration standard curves (test vs reference), comparing QC samples to an international reference standard or by analyzing incurred study samples using two different calibration standards (test vs reference). Four different normalization concepts are listed in [Table 1](#) and are described in the following paragraphs. In each case after normalization, the dilution scheme of the new lot of the calibration material should be adjusted accordingly in order to keep the original validated LLOQ and ULOQ of the assay unchanged.

Whenever possible, statistical evaluation for meaningful differences between the test and reference data should be made. F- and t-tests are performed for comparing the variances and means of the two distributions. If in the F-test the SDs are not statistically different, then the two-sided t-test can be applied to test the significance of the means. Only if the means of the tested data populations are shown to be statistically different then one of the following approaches can be applied for normalizing the test material. An alternative statistical approach would be the equivalence test, which is not particularly dependent on the variance of the actual measured results. Other approaches such as normalizing in every situation or normalizing only when the difference is greater than two- to three-times the intra-assay or interassay precision may be justified depending on the use of the biomarker assay. It is necessary to describe the normalization concept and the criteria when to normalize upfront in the analytical work plan or the method description.

For large number of test samples, multiple experiments and well defined statistics may be preferable [5]; however, we also focus on more simple concepts that can be performed without huge effort and a simple calculation-based worksheet.

Comparing test with reference samples

This concept is recommended for the normalization of test versus international reference standard material. Normalization requires checking the previous QC concentration results if the deviation from the target is independent from the concentration level (parallelism test).

Samples

A test sample (mid range) is prepared from the new lot of calibration standard material in the sample matrix

(usually serum/plasma) or calibration sample matrix.

A reference sample (at the same level as for the test sample) is prepared from the original lot of calibration standard material or international reference standard in the same matrix.

Experiment

Test and reference samples are measured in minimum of five independent replicates (independent dilutions) on 1–3 assay runs (depending on the inter-assay pre-

Table 1. Various concepts to normalize concentration of standard.

Normalization concepts	Samples	Experiments	Evaluation	Recommendation
A Test versus reference calibrator sample	<ul style="list-style-type: none"> Test sample in calibrator matrix, prepared from new lot calibration standard (mid of calibration range) Reference sample in calibrator matrix, prepared from original lot calibration standard or reference/ international standard (same level as test sample) 	Measurement of test and reference sample in three to five independent replicates (independent dilutions) on 2–3 days/ runs (n = 6–15) together with original QC set versus new calibration samples (prepared from new lot)	<ul style="list-style-type: none"> New concentration (new lot) = mean test/ mean reference * label concentration (new lot) If % CV of QC results (n = 6–15) ≤ precision of assay and QC results are parallel to calibration curve and mean test sample is significant different from mean reference sample (e.g., by equivalence test) 	Recommended for in-house developed assays to check change of calibration standard lot, only possible approach for normalization versus international standard
B Matrix samples versus calibration curves	Incurred matrix samples as test samples (e.g., QCs original and/or new, monitoring samples/ sample controls)	Measurement versus calibration samples prepared from original lot 1 and versus calibration samples prepared from lot 2 (multiple batches and replicates)	New concentration (new lot) = mean result (vs calibration samples from lot 2)/mean results (vs calibration samples from lot 1) * label concentration (new lot)	Recommended if not only lot of calibration standard changes but also kit lot which may also include a change of lots of critical reagents
C New versus original calibration curve	Test calibration samples (whole curve range) from new lot, reference calibration samples (curve) from original lot	Based on performance of calibration samples. Read out with both new and original calibration samples. New calibrations samples measured versus original calibration samples	<ul style="list-style-type: none"> New concentration (new lot) = label concentration (new lot)/ mean recovery new calibration samples, If calibration curves are parallel Confirmed by equivalence test with calibration samples between runs with original standard and runs with new standard 	Recommended when no international standard or other external standard is available
D Linear correlation	30 or more incurred samples (e.g., remaining study samples)	Measurement of samples with original and new lot (calibration standard or kit)	Linear correlation of results (new lot vs original lot). 1/slope is used as normalization factor to adjust new calibration standard/kit	

QC: Quality Control.

cision) together with previous set of QC samples (low, medium and high). Calibration standards are prepared from the new lot of calibration standard.

Evaluation

Equation 1 can be applied for adjusting the concentration of the new lot.

$$\text{New concentration (new lot)} = (\text{mean test value} / \text{mean reference value}) \times \text{label concentration (new lot)} \quad (1)$$

An example of comparing the new lot of reference standard to the WHO reference standard is shown in Table 2. Both standards were diluted to a value of 140 ng/mL and analyzed against the calibration curve prepared from the newly prepared lot.

Based on the experimental data, the nominal value of the new standard lot label concentration was adjusted from 1.00 to 0.744 mg/ml (137/184).

After adjusting the concentration of the new standard material the previous set of QC samples (low, medium and high) prepared from a previous lot should show acceptable results when analyzed against the new standards.

Comparing QC samples against calibration curves prepared from original & new lots

This concept is recommended when no international standard or other external standard is available. It can be applied also when comparing commercial kit lots or their critical reagent lots. Again, proof of parallelism is a prerequisite for normalization. This information is available from the QC results and the variability of the normalization factor (NF) calculated for each of the three QCs.

Samples

QC samples (n = 3, low, mid and high) are used as test samples. Calibration samples (whole curve range) are prepared from the new and original lots of calibration standard material.

Experiment

QC samples are measured against calibration samples prepared from the new (test sample) and original (reference sample) lots of calibration standard materials on 1–3 assay runs (depending on the inter-assay precision).

Evaluation

Equation 2 can be applied for adjusting the concentration of the new lot.

$$\text{NF} = \frac{\text{Mean result vs. calibrator (new lot)}}{\text{Mean results vs. calibrator (original lot)}} \quad (2)$$

The new concentration (new lot) will be the mean result of NFs from each QC sample multiplied by the label concentration of the new lot (the % CV of the mean results should not exceed a predefined acceptance criteria, e.g., 20–30%).

An example of comparing the new lot of reference standard to the original lot of reference standard is shown in Table 3. Three levels of QC samples (each in three replicates) were analyzed against calibration curves prepared from the new and original lots.

Comparing calibration curves

This concept also controls the parallelism between the new and original calibrator standard materials without a need to run additional QC samples.

Table 2. Example for normalization concept A.

Nominal value	Code	Concentration	Unit	% deviation from nominal	Mean concentration	% CV	NF
140	WHO_1	187	ng/mL	34			
140	WHO_2	179	ng/mL	28			
140	WHO_3	187	ng/mL	34			
140	WHO_4	190	ng/mL	36			
140	WHO_5	178	ng/mL	27	184	2.9%	
140	STD_1	136	ng/mL	-3			
140	STD_2	129	ng/mL	-8			
140	STD_3	145	ng/mL	4			
140	STD_4	135	ng/mL	-4			
140	STD_5	140	ng/mL	0	137	4.3%	0.744
New conc. = $\frac{\text{Mean conc. (NEW STD LOT)}}{\text{Mean conc. (WHO REF STD)}} \times \text{New lot nominal STD conc.} = \frac{137 \text{ ng/mL}}{184 \text{ ng/mL}} \times 1.00 \text{ mg/mL} = 0.744 \text{ mg/mL}$							
NF: Normalization factor.							

Table 3. Example for normalization concept B.

Nominal value	Code	Concentration	Unit	% deviation from nominal	Mean concentration	NF (new/old)	% CV (NF)	Mean NF
20	LOW-QC_OLD STD_1	19.6	ng/mL	-2				
20	LOW-QC_OLD STD_1	16	ng/mL	-20				
20	LOW-QC_OLD STD_1	18.3	ng/mL	-9	18.0			
20	LOW-QC_NEW STD_1	33.6	ng/mL	68				
20	LOW-QC_NEW STD_1	29.3	ng/mL	47				
20	LOW-QC_NEW STD_1	30.9	ng/mL	55	31.3	1.74		
100	MID-QC_OLD STD_1	98.4	ng/mL	-2				
100	MID-QC_OLD STD_1	87.9	ng/mL	-12				
100	MID-QC_OLD STD_1	96.6	ng/mL	-3	94.3			
100	MID-QC_NEW STD_1	145	ng/mL	45				
100	MID-QC_NEW STD_1	133	ng/mL	33				
100	MID-QC_NEW STD_1	138	ng/mL	38	138.7	1.47		
300	HIGH-QC_OLD STD_1	298	ng/mL	-1				
300	HIGH-QC_OLD STD_1	285	ng/mL	-5				
300	HIGH-QC_OLD STD_1	275	ng/mL	-8	286.0			
300	HIGH-QC_NEW STD_1	423	ng/mL	41				
300	HIGH-QC_NEW STD_1	445	ng/mL	48				
300	HIGH-QC_NEW STD_1	431	ng/mL	44	433.0	1.51	9.2%	1.575

New concentration = mean NF × new lot nominal standard concentration (1.575 × 1.00 mg/mL = 1.575 mg/mL).
NF: Normalization factor.

Samples

Test calibration samples (whole curve range) are prepared from the new lot of calibration standard material. Reference calibration samples (whole curve range) are prepared from the original lot of calibration standard material.

Experiment

Calibration samples prepared from the new lot (test sample) are measured against calibration samples prepared from the original lot (reference sample) on 1–3 assay runs (depending on the interassay precision).

Evaluation

Equation 3 can be applied for adjusting the concentration of the new lot:

$$\text{NF} = \frac{\text{Analysed result of the calibrator sample within measurement range (new lot)}}{\text{Calibrator label concentration (new lot)}} \quad (3)$$

The new concentration (new lot) will be the mean result of NFs from each calibration sample level multiplied by the label concentration of the new lot (the % CV of the mean results should not exceed a predefined acceptance criteria, e.g., 20–30%).

An example of comparing the new lot of reference standard to the original lot of reference standard is shown in Table 4. Calibration samples were prepared

from the new and original lots and the new lot was analyzed against the original lot.

Comparing incurred matrix samples against new & original calibration curves

This concept is recommended when no international standard or other external standard is available. It can be applied also when comparing commercial kit lots or their critical reagent lots. When incurred matrix samples span the entire dynamic range of the assay, this approach also controls the parallelism between the new and original calibration standard materials without a need to run additional QC samples.

Samples

Incurred matrix samples ($n \geq 30$) are used as test and reference samples. Calibration samples (whole curve range) are prepared from the new and original lots of calibration standard material.

Experiment

Incurred matrix samples are measured against calibration samples prepared from the new (test sample) and original (reference sample) lots of calibration standard materials on 1–3 assay runs (depending on the interassay precision).

Evaluation

Equation 4 can be applied for adjusting the concentration of the new lot.

$$\text{NF} = \frac{\text{Mean result vs. calibrator (new lot)}}{\text{Mean result vs. calibrator (original lot)}} \quad (4)$$

The new concentration (new lot) will be the mean result of NFs from each incurred matrix sample multiplied by the label concentration of the new lot (the % CV of the mean results should not exceed a predefined acceptance criteria, e.g., 20–30%).

Alternatively, linear correlation of results of incurred matrix samples measured versus new and original lots can be used. The NF (1/slope) is used to adjust the new calibration standard material, as in Equation 5.

$$\text{New concentration (new lot)} = (1/\text{slope}) * \text{label concentration (new lot)} \quad (5)$$

An example of using in current samples is shown in Figure 1.

Check for parallelism of the two different calibration standard lots

A prerequisite of any concept for normalization of calibration standards is a bias between immunoreactivity of both lots that is independent from the concentration of analyte.

This kind of parallelism can be tested between the two calibration curves prepared from the original and the new (or the new and the reference) lot (concept C), the calibration curve prepared from the new lot and diluted incurred matrix samples (QCs or sample controls, concept A and B) or between sample results measured versus calibration curves prepared from both lots of calibration standard. The latter case is used in normalization concept D. A nonlinear correlation would indicate nonparallelism between lots.

The ideal result would be for parallelism to be unaffected by a lot change but this is not always the case.

Depending on at which point on the curve the nonparallelism occurs it may be acceptable to truncate the dynamic range of the assay and normalize results for the area of the curve where parallelism is observed. For example, if nonparallelism is demonstrated toward the lower concentrations of the curve and the tested samples are not expected to fall within this range then this approach could be applied. However, one caveat would be that a sufficient number of calibrator concentrations remain in the parallel range of the assay curve. Additionally there is a risk that a lower number of calibration samples would result in fewer reported concentrations when using a truncated assay range.

For cases where nonparallelism is seen across a significant portion of the curve, normalization would not be a suitable approach to deal with the lot-to-lot change as the normalization factor NF becomes dependent on the analyte concentration, which would require a normalization function instead of a constant factor. Instead, it is highly likely that a significant change in the assay and a revalidation would be required to address both the nonparallelism and the lot change.

In addition to parallelism of the two calibration standard lots, as a prerequisite for normalization by a concentration independent factor, it might be pertinent to check the assay performance for general parallelism [12], for example, after changes of the source of calibration standard (different vendor, different cell line and different isoform) or after a change of the assay kit (e.g., different vendor). This would then be a form of partial revalidation of the assay.

Trend analysis

Another tool to take care of the performance of the method and the comparability of results measured over a long period of time (even across changes of reagent lots) is trending. Trending of results for biomarker assays is already used by many to assess QC stability and assay robustness [13].

Table 4. Example for normalization concept C.

Nominal value	Code	Concentration	Unit	% deviation from nominal	NF (NEW/OLD)	% CV (NF)	Mean NF
10	NEW STD_1	<LLOQ	ng/mL	–	–		
20	NEW STD_2	15.2	ng/mL	-24	0.760		
40	NEW STD_3	30.9	ng/mL	-23	0.773		
80	NEW STD_4	64.8	ng/mL	-19	0.810		
160	NEW STD_5	128	ng/mL	-20	0.800		
320	NEW STD_6	251	ng/mL	-22	0.784		
640	NEW STD_7	540	ng/mL	-16	0.844	3.8%	0.795
New concentration = Mean NF × New lot nominal standard concentration (0.795 × 1.00 mg/mL = 0.795 mg/mL). NF: Normalization factor.

Figure 1. Example for normalization concept D. Comparing incurred matrix samples measured against new and old calibration curves by linear regression analysis. The slope of 0.862 was obtained from the linear regression analysis. In this case, the original label concentration of the new lot would be multiplied by a factor of 1.16 ($1/0.862$).

QCs serve a specific purpose during assay validation of determining accuracy and precision, yet during sample analysis they are employed for assay run acceptance or rejection using, for example, the 4–6–20 rule as described in the conference report of the 3rd AAPS/FDA Bioanalytical Workshop [14]. However, it is important to consider the performance of an assay over time rather than in isolated sample runs, not just for single reagent and kit lots, but also across lots. Both the EMA guideline [15] and the US FDA guidance on bioanalytical method validation [16] as well as the proficiency testing criteria set out in the Clinical Laboratory Improvement Amendments require evaluation to determine that the performance characteristics of the bioanalytical method are not altered when the calibration standard is changed. When lot changes occur, QCs should be employed to assess the impact of the change on the assay method using predefined criteria, preferably comparing both the original and new lots.

In-study assay performance monitoring through QC trend analysis is equally important. However, for biomarker assays it is often the case that the QCs may

not accurately reflect the study samples due to the use of a recombinant form of the analyte for preparation of the QCs. For many biomarkers, there is an appreciable endogenous concentration in readily available commercial samples. Therefore, it is recommended that a 'matrix QC' is created from such a sample, or pool of samples, and routinely included in sample analysis. Additionally it may be possible to produce several matrix QCs that span the dynamic range of the standard curve. Inclusion of such samples affords confidence that comparable results are produced each time the assay is utilized and would detect upward or downward trends in concentration assuming that stability of such samples can be determined. Such an approach is described by Wang *et al.* [13], where both high and low 'authentic samples' (incurred samples) were analyzed in each sample analysis run. Furthermore they were able to demonstrate reliability and robustness for the method despite two different reference material lots being used over 3 years in 12 clinical studies.

While QCs are important in trend analysis, other components of the assay should not be overlooked.

Assay variability and reproducibility can also be detected through the monitoring of parameters such as the signal response of the zero analyte or blank, the maximal response of the standard curve or highest calibrator as well as the slope of the calibration curve. It is therefore recommended that these parameters are recorded and documented during sample analysis along with the concentration of QC samples. Additionally it is prudent to include the lot numbers of the assay reagents or the lot number of the kit.

There are various trend analysis options that can be employed depending on the decision making utility of the biomarker:

- Simple graphical depiction of QC concentration over time against different lots of calibrator and manually monitoring upward or downward trends; a basic approach to trigger further investigation;
- Approach 1 plus the incorporation of an acceptance limit such as a 20% deviation from a predefined mean concentration (as depicted in Figure 3);
- Approach 1 with the addition of statistical tools such as Levey-Jennings, utilizing a 2 or 3 SD of the mean, or employing a multirule QC approach such as Westgard rules;
- Monitoring the distributions of healthy volunteers or actual subject results in an adequately sized population and determining a median and a range value, assuming that biological variability is not excessively large; a thorough description of this approach is discussed in Algeciras-Schimmich *et al.* [17].

Whichever approach is taken, it is recommended that the timing, the method of documentation and statistical approach (if applicable) is determined *a priori*.

Perspectives for preferred strategies how to deal with variable calibration standards

When performing biomarker analyzes that require the use of several lots of calibration standards, there are multiple strategies that can be considered. Each approach has its own advantages, limitations, complexities and

Figure 2. Decision tree for choice of an appropriate approach regarding handling lot-to-lot variability of calibration standard.

future implications; discussions of which have been the focus of this EBF team. The broad range of intended uses of biomarker measurements performed during drug development requires a tiered approach not only for assay validation but also regarding effort that is spent on the comparability of results [4,18]. As a result of several discussions and a survey performed within the EBF community it became obvious that various strategies are currently in use and that there is no single strategy that is applicable to all cases. Nevertheless, the EBF would like to recommend some general principles on how to determine an acceptable approach:

- Apply the scientific tiered approach of biomarker assay validation. Use the qualification status of the Biomarker (exploratory/confirmatory), the intended use of the data, as well as the validation level of the assay as the basis for the decision on the optimal approach for dealing with lot-to-lot changes of calibration standards;
- Define the maximum acceptable method bias originating from changing calibration standard lots in the upcoming clinical or nonclinical study. As biomarker assays are often developed and validated according to the 'Fit-For-Purpose' strategy, the maximum acceptable bias will depend on intended purpose of the measurements, biological variability and assay performance;
- During method development, initiate a process for how to assess lot-to-lot variability of calibration standards. Several lots from each supplier as well as different suppliers should be included in the assessment and the choice of calibration standard supplier should be based on consistency between lots;
- The availability of sufficient amounts of calibration standard, reference standard or an international standard would be another factor when choosing a strategy. This is especially important when the use of commercial assay kits with limited amounts of reagents may limit the ability to bank large quantities of the calibration standard;
- Select any of following options depending on biomarker status, validation level and knowledge about lot-to-lot variability:

Exploratory biomarkers option 1 (basic method validation level)

This approach, which corresponds to approach #1 (Table 5), is recommended for exploratory biomarkers and a basic validation level of the assay without the need of a predefined lot change strategy (e.g. quasi-quantitative or relative changes in data). It applies to the use of poten-

Figure 3. Case study: monitoring of quality control performance (trending). Quality controls failed after the switch to a new calibration standard lot B as they exceeded acceptance limits. Using a normalization concept runs measured versus lot B would have passed.

tially expensive commercial assay kits when minimal effort should be spent upfront without knowing whether the biomarker of interest meets expectations or not. Measurement of study samples will be initiated with one lot of kits. If the study lasts longer than the availability of the first kit lot then any subsequent lot would be used without any bridging experiments, but the QCs would be retained for trend analysis and comparison of run performance. The switch from the original to the new kit lot should be performed on the same day if possible, to minimize the parameters that might influence assay performance. Overall, performance of all kits lots and comparability of results will be checked retrospectively based on results of QC samples. Depending on the overall assay precision, the biological variability and the observed change in biomarker level, the results can be kept as measured without any adjustment (but with potentially higher variability). If QC results measured with one lot differ significantly from the QC results measured with the other lots, all sample results measured with the new lot could be adjusted retrospectively by an NF, for example calculated according to normalization concept B (Table 1).

Regardless of whether lot-to-lot variability has been investigated, a key risk to this approach is the generation of less or even noncomparable results within or between studies.

Exploratory biomarkers option 2 (advanced method validation level), & low lot-to-lot variability &/or short-lasting studies

Under these conditions it is our recommended approach to use one lot within the expiry date and to change the lot at expiry date without doing any normalization. Due to the short term expiry dates applied

Table 5. Experimental approaches to guarantee long-term comparability of calibration standards.

No.	Tool/approach	Advantages	Challenges, limitations	Suggested for validation levels	Requirements	Normalization/trending
1	<ul style="list-style-type: none"> Retrospective risk analysis and management Use several lots of calibration standard without any <i>a priori</i> comparison or correction strategy 	<ul style="list-style-type: none"> No additional work required 	<ul style="list-style-type: none"> Not science based Risk of noncomparable results Introduction of bias 	<ul style="list-style-type: none"> Exploratory biomarker Basic validation level 	<ul style="list-style-type: none"> None 	<ul style="list-style-type: none"> Trending of QC results may help support this approach
2	<ul style="list-style-type: none"> Use of one lot only within expiry date Use of one lot of calibration standard within a particular project or study within expiry date of vendor 	<ul style="list-style-type: none"> No risk of lot-to-lot variability No additional work required 	<ul style="list-style-type: none"> Only suitable for short studies May be difficult to compare results across different studies where different lot numbers have been used Risk of study extension resulting in calibration standard falling outside of expiry 	<ul style="list-style-type: none"> Exploratory biomarker Basic or advanced validation level 	<ul style="list-style-type: none"> Purchase sufficient material at right time Careful timing of order and sample measurement 	<ul style="list-style-type: none"> Not applicable
3	<ul style="list-style-type: none"> Use of one lot only ignoring expiry date Banking (at vendor or in-house) and use of one lot of calibration standard across numerous studies Storage of standards at $\leq -80^{\circ}\text{C}$ or in lyophilized form ignoring the expiry date from the vendor 	<ul style="list-style-type: none"> No risk of lot-to-lot variability 	<ul style="list-style-type: none"> High frontloading costs For commercial immunoassay kits it may only be possible if a separate calibration standard is available Requires assumption of long-term stability 	<ul style="list-style-type: none"> Exploratory biomarker, advanced validation level 	<ul style="list-style-type: none"> Purchase sufficient material, store in aliquots Monitoring of assay performance (sample or QC – analyte concentration or assay response) 	<ul style="list-style-type: none"> May be beneficial to check results against reference population range Trend analysis of banked monitoring samples
4	<ul style="list-style-type: none"> Use of multiple lots with normalization to the previous lot 	<ul style="list-style-type: none"> No banking of reference standard required Commercially supplied kit material can be used 	<ul style="list-style-type: none"> Long-term trends may be ignored Requires multiple bridging experiments and normalization strategy with available standard from previous lot Considerable work should there be several lot changes over course of study/project 	<ul style="list-style-type: none"> Exploratory or confirmatory biomarker, advanced or full validation level 	<ul style="list-style-type: none"> Performance of several batches using both new and original lots of calibration standard Possible inclusion of independent quality controls or incurred samples 	<ul style="list-style-type: none"> Normalization concept required – suggested B (Table 1) Trending of QC data subsequent to normalization is recommended

QC: Quality control.

Table 5. Experimental approaches to guarantee long-term comparability of calibration standards (cont.).

No.	Tool/approach	Advantages	Challenges, limitations	Suggested for validation levels	Requirements	Normalization/trending
5	Use of multiple lots with normalization to a banked reference standard lot or international standard	<ul style="list-style-type: none"> Kit material can be used 	<ul style="list-style-type: none"> Possible long-term stability issues with banked reference standard Requires multiple bridging experiments and normalization strategy which takes into account whether it is possible to prepare a full calibration curve from an international standard 	<ul style="list-style-type: none"> Confirmatory biomarker, full validation level 	<ul style="list-style-type: none"> Performance of several batches using both new and original lots of calibration standard and international standard Possible inclusion of independent quality controls or incurred samples 	<ul style="list-style-type: none"> Normalization concept required – suggested A (Table 1) Trending of QC data subsequent to normalization is recommended

QC: Quality control.

to commercially supplied calibration standards or kits, this approach may only be suitable for short-term clinical biomarker studies for example Phase I. It may also be a useful strategy where lot-to-lot differences have previously been demonstrated (e.g., in method development) as being low or not critical for the intended use of the assay. In these instances sufficient quantities of calibration standard (of a suitable expiry date) should be obtained to cover the entire analysis. This approach corresponds to approach #2 (Table 5).

Exploratory biomarkers option 2 (advanced method validation level) & risk for significant lot-to-lot variability & availability of a bulk calibration standard

Under these conditions the more laborious and complex normalization approaches can be avoided by using one lot of calibration standard only, stored under extraordinary or optimized conditions, beyond the expiry date. Examples of such storage conditions are storage in lyophilized form or storage at ultra-low temperatures. As no regulatory requirement is as yet defined for an exploratory biomarker there is no formal requirement on validation of the stability of the reference standard covering the full period of time. However, since the data from monitoring of an explorative biomarker might be used for important internal decision-making, it is strongly recommended that this approach is used with careful monitoring of assay performance by trend analysis assessment using incurred samples. This approach corresponds to approach #3 (Table 5).

A key assumption is that the stability of the standard is assured beyond the supplied expiry date and that sufficient material can be purchased up-front from the vendor. In cases where the calibration standard is a kit component, this may lead to wastage of other reagents (e.g., antibodies or coated plates). A viable alternative would be to use a calibration standard from a different manufacturer as a substitute.

If sufficient bulk quantities of calibration standard are not available (e.g., in case of assay kits) the following approach number 4 can be chosen.

Exploratory or confirmatory biomarkers & no availability of an international standard or banked reference standard

Under these conditions it is necessary that the same biomarker assay will be deployed over a longer time frame, and that lot changes in the calibration standard will inevitably occur; however, the analyst needs to compare data without bias. This also presents an alternative approach for commercially supplied kits, removing the requirement to use expired reagents or alterna-

tive calibration standards. In essence each new lot of calibration standard is normalized against the previous lot to remove any bias in the data. Potential pitfalls arise where previous lots are no longer available, for example, if an assay is utilized after a significant break, or where a previous lot has fallen outside of the expiry date. Another drawback would be an undetected long-term drift of the assay through multiple lot changes. This would result in a large bias when comparing data from lot one with those from a later lot. Therefore it is recommended to store sample controls or other comparator samples and to perform a trending approach in order to monitor long-term drift of the assay. This approach corresponds to approach #4 (Table 5).

Confirmatory biomarkers & availability to either an international or a banked long-term stable reference standard

Under these conditions normalizing each lot of calibration standard to an international or banked reference standard is the recommended choice (e.g., using normalization concept A, Table 1). The international standard or in-house established reference standard would be part of a global reference system that guarantees proven metrological traceability and comparability of results measured in various laboratories over long time with the same assay. Such an approach is also used for *in vitro* diagnostic assays. It is an optimal strategy for fully validated biomarker assays where quantitative biomarker data are being compared or pooled, for example, for meta analysis across various pivotal trials. This approach corresponds to approach #5 (Table 5).

For both approaches 4 and 5, a predefined procedure describing how to deal with the lot-to-lot variability, including when and how to normalize calibration standards, should be determined. A preset acceptance criterion should be defined based on assay performance and intended use of the assay. A decision tree for selecting an appropriate calibration standard approach is shown in Figure 2.

Conclusion

This paper provides the outcome of an intense EBF internal discussion, including a survey to all members involved in protein biomarker bioanalysis using LBAs. In principle the discussed views and perspectives are equally relevant for other quantitative analytical techniques and analytes, whenever the reproducible production of the calibration standard may be a challenge.

The experience and the presented case studies clearly indicate the importance of taking control of lot-to-lot variability of calibration standards, critical reagents

and assay kits in order to guarantee comparable results over long-term of clinical studies or even the whole drug development period. The use of research grade assay kits without defining comparator samples or reagents for run-to-run comparisons, together with the lack of knowledge about the biology of the biomarker and the possible influence of the sampling procedure on the results, are the key factors that could result in noninterpretable or even misinterpreted results; the ultimate outcome being a waste of human biospecimens and resources within clinical study centers and bioanalytical laboratories.

Future perspective

As an industry, commercially available biomarker kits are often utilized to support clinical trials and internal decision-making. However, these kits are generally not validated by the vendor and are marked as 'for research use only'. This can produce several challenges especially related to the control and consistency of the kit components. Additionally the antibodies within the kits may not be fully characterized and in some cases the reagents are not specific or selective and may demonstrate binding to other analytes. Equally there can be issues with the calibration standards used in such kits.

Going forward, the bioanalytical community would welcome fully standardized and well controlled critical reagents to ensure stability and consistency across different reagent lots. Such reagents and kits would undergo improved characterization and commercial vendors would be open to sharing more comprehensive information on their QC processes, how they determine the suitability for use of each reagent and provide fully comprehensive CoFA documentation. Ideally where manufacturers use commercial antibodies within their biomarkers kits, they would consider producing their own reagents to provide greater control of consistency and potentially longer expiry dates.

Additionally when lot changes are inevitable, it would be beneficial if commercial suppliers ensured timely communication to end users to allow assessment of the impact and provide adequate time to take any required mitigation steps.

Disclaimer

The views and conclusion presented in this paper are those of the European Bioanalysis Forum and do not necessarily reflect the representative affiliation or company's position on the subject.

Acknowledgements

The authors would like to thank the EBF-IGM members for their contributions and impact to our topic team.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employ-

ment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Executive summary

- The EBF recommends to choose the calibration standard of a biomarker assay or an assay kit carefully with regard to the suitability for the intended use (and whether it is representative of the endogenous analyte), availability of sufficient amounts over long-term and the constant quality of the standard over several lots.
- The EBF recommends that long-term use of protein biomarker assays always require an approach for dealing with lot-to-lot variability of calibration standards. The minimum strategy would be simply monitoring assay performance and performing trending analysis, the maximum being referencing and normalizing calibration standards against a qualified reference or international standard.
- The EBF recommends that the approach is chosen based on the qualification level of the biomarker, the intended use of the biomarker data, the method validation level, the knowledge about lot-to-lot variability and the availability of standards.
- The EBF recommends that the chosen approach should be documented *a priori* in relevant documents such as the method description, the validation plan, the analytical work plan of a study or in a standard operating procedure.
- The EBF recommends to discuss the requirements for biomarker calibration standards in the context of long-lasting drug development with the vendors of such reagents and assay kits to ultimately provide a high standard and reproducible product and to compel them to optimize their production and quality control procedures.

References

- 1 Wild D, Sheehan C. Standardization and calibration. In: *The Immunoassay Handbook (4th Edition)*. Wild D (Ed.). Elsevier, Oxford, UK, 315–322 (2013).
- 2 Sweep FCGJ, Fritsche HA, Gion M *et al.* Considerations on development, validation, application and quality control of immuno(metric) biomarker assays in clinical cancer research: an EORTC-NCI working group report. *Int. J. Oncol.* 23, 1715–1726 (2003).
- 3 Cull CA, Manley SE, Stratton IM *et al.* Approach to maintain comparability of biochemical data during long-term clinical trials. *Clin. Chem.* 43(10), 1913–1918 (1997).
- 4 Timmerman P, Herling C, Stoellner D *et al.* European Bioanalysis Forum recommendation on method establishment and bioanalysis of biomarkers in support of drug development. *Bioanalysis* 4(15), 1883–1894 (2012).
- 5 Jani D, Allinson J, Berisha F *et al.* Recommendations for use and fit-for-purpose validation of biomarker multiplex ligand binding assays in drug development. *AAPS J.* 18(1), 1–14 (2016).
- 6 Bower JF, McClung JB, Watson C *et al.* Recommendations and best practices for reference standards and reagents used in bioanalytical method validation. *AAPS J.* 16(2), 352–356 (2014).
- 7 Lee JW. Method validation and application of protein biomarkers: basic similarities and differences from biotherapeutics. *Bioanalysis* 1(8), 1461–1474 (2009).
- 8 Lee JW, Devanarayan V, Barrett YC *et al.* Fit-for-purpose method development and validation for successful biomarker measurement. *Pharm. Res.* 23(2), 312–328 (2006).
- 9 Thway TM, DeSilva B. Stability assessment in ligand-binding assays: a critical parameter for data integrity. *Bioanalysis* 7(11), 1315–1317 (2015).
- 10 Amaravadi L, Song A, Myler H *et al.* 2015 White Paper on recent issues in bioanalysis: focus on new technologies and biomarkers (Part 3 – LBA, biomarkers and immunogenicity). *Bioanalysis* 7(24), 3107–3124 (2015).
- 11 Nicholson R, Lowers S, Caturia MC *et al.* 6th GCC focus on LBA: critical reagents, positive controls and reference standards; specificity for endogenous compounds; biomarkers; biosimilars. *Bioanalysis* 4(19), 2335–2342 (2012).
- 12 Stevenson LF, Purushothama S. Parallelism: considerations for the development, validation and implementation of PK and biomarker ligand-binding assays. *Bioanalysis* 6(2), 185–198 (2014).
- 13 Wang J, Lee J, Burns D *et al.* “Fit-for-purpose” method validation and application of a biomarker (C-terminal telopeptides of type I collagen) in denosumab clinical studies. *AAPS J.* 11(2), 385–394 (2009).
- 14 Viswaanathan CT, Bansal S, Booth B *et al.* Workshop/conference report – quantitative bioanalytical methods validation and implementation: best practices for chromatographic and ligand binding assays. *AAPS J.* 9(1), E30–E42 (2007).
- 15 European Medicines Agency, Committee for Medicinal Products for Human Use. *Guideline on Bioanalytical Method Validation*. London, UK (2011). http://www.ema.europa.eu/docs/en_GB/document_library
- 16 US Department of Health and Human Services, US FDA, Center for Drug Evaluation and Research, Center for

- Veterinary Medicine. *Guidance for Industry, Bioanalytical Method Validation*. MD, USA (2001).
<https://www.fda.gov/downloads/Drugs/>
- 17 Algeciras-Schimmich A, Bruns DE, Boyd JC *et al*. Failure of current laboratory protocols to detect lot-to-lot reagent differences: findings and possible solutions. *Clin. Chem.* 59(8), 1187–1194 (2013).
- 18 Timmerman P. Tiered approach revisited: introducing stage-appropriate or assay-appropriate scientific validation. *Bioanalysis* 6(5), 599–604 (2014).