GRANT AGREEMENT # 115003

Safe-T

Safer and Faster Evidenced-Based Translation

STANDARD VALIDATION PROCEDURE FOR BIOMARKER IMMUNOASSAYS
Version 4.1

Authors:

Jens Göpfert (NMI)
Kaïdre Bendjama (Firalis)
Thomas Knorpp (EDI/NMI)
Max Breitner (Firalis)
Thomas Schindler (Roche)
Stefanie Rimmele (NMI)
Mark Pinches (AstraZeneca)
Nicole Schneiderhan-Marra (NMI)
Barbara Fischer (Firalis)
Rory Connolly (EKF Diagnostics (former: Argutus Medical))
Håkan Andersson (AstraZeneca)

Reviewers:

Thomas Joos (NMI/EDI/MRBM)
Joe Keenan (EKF Diagnostics (former: Argutus Medical))
Marie-Helene Pascual (Sanofi-Aventis)
Joachim Stangier (Boehringer Ingelheim)

PART A – STANDARD IMMUNOASSAY VALIDATION PROCEDURE WITHIN IMI SAFE-T

A.1 Scope

The procedures set out in this SOP apply primarily to enzymatic and fluorescent immunoassays which are used for the quantitative determination of biomarkers in samples from human origin within the SAFE-T project. The procedures for other assay types such as LC-MS or qPCR should adapt the proposed procedures if applicable.
It should be noted that the experiments outlined in this document represent best practice as defined by the IMI SAFE-T WP5. WP5 recognizes that all immunoassays may not be able to achieve the quality standards described, or that it may not be appropriate for technical reasons to perform certain experiments for specific assays or assay types. However it is expected that the reasons why specific experiments cannot be performed, or why specific quality standards are not met, are described in the immunoassay validation report.

A.2 Introduction

Validation is necessary to demonstrate the performance and the reliability of the assay. Method validation is governed by various guidelines edited by regulatory bodies (e.g European Medicine Agency, 2009; 2001 and Food and Drug Administration 2013) or institutes defining clinical and laboratory standards (e.g. CLSI) for generation of data in human samples. However, many existing guidelines are biased towards chromatographic methods for the detection of drugs in human samples in PK or bioequivalence studies and specific guidance for ligand binding assays were only released very recently mentioning the data generation to support biomarker qualification (FDA 2013). Herein we describe a procedure for the validation of immunoassays for data generation to support biomarker qualification. This SOP is derived from the assay validation policy of IMI SAFE-T. The assay validation process is split into several phases with increasing levels of validation (pre-exploratory, exploratory, and pre-confirmatory) according to the "fit-for-purpose" approach (Lee et al., 2006; Lee & Hall, 2009).

A.3 Operating instructions

Pre-Exploratory phase

Required experiments in the pre-exploratory phase (for "low-bar validation"):

- **Calibration curve validation** ensures that the curve fitting model used for the calculation of the standard curve is adapted to the assay and provides a robust

description of the concentration-response of the assay. To do so back-calculated concentration of standards used during the pre-exploratory validation are pooled and the residuals are plotted against the concentration of standard. The variance needs to be homogenous (see further description in section B.8.1.).

- **Recovery**: Spiked sample recovery studies must be evaluated using an appropriate matrix. Investigations are run primarily as an indication of sample matrix effects. Recovery evaluations will be made by using human serum, plasma and/or urine sample with no or low analyte concentration. Three samples will be spiked using a minimum of three different concentrations of the standard curve reagent (low, mid, high). Acceptable performance will be based on percent recovery at each spiked concentration. Perform triplicates.

  Recovery should typically be within 80-120% over the working range of the assay. Lower or higher level of recovery may be acceptable if it is maintained over the working range of the assay.
- **Linearity/Reportable range**. Use plasma, serum or urine with a measureable concentration (preferably in the higher part of the working range), dilute with matrix, water, saline, or supplied sample diluent (whichever is most appropriate). The assay should be linear over the range of dilutions which may be necessary to measure concentration in patient samples. Dilutions of the sample matrix shall be selected to cover this range. Typically observed value should be within 80-120% of the expected value. Perform triplicate.
- **Precision: Intra-assay.** To establish reproducibility of an assay by comparing replicates of a sample with a measurable concentration (preferably in the middle of the range) within one run on the same day, expressed as mean, standard deviation and % CV. Use most appropriate matrix possible (plasma/serum/urine pool > synthetic matrix>standard solutions or buffer). N=10 minimum (preferably N=20). The kit QCs (if available) must also produce acceptable results (N=10 minimum). Intra-assay precision should not exceed a CV of 15%.
- **Precision: Inter-assay.** To establish reproducibility of an assay by comparing triplicate measurements of each sample per run of independent runs on 3 different days, expressed as mean, standard deviation and % CV. Samples should span the range of the standard curve (best to have 2 in the low, 1 in the middle and 2 in the high concentration range). Inter-assay precision should not exceed a CV of 20% for each point (30% for the low concentration samples).
- **Lower limit of detection (LOD)**. Lowest concentration of an analyte that is significantly different from a blank sample + three standard deviations (SD). Analysis of at least 20 replicates of the zero calibrator, analyte free serum /plasma/urine or equivalent material will be performed. The mean raw data signal of the zero and next lowest calibrators are calculated and a line fit is used. Depending on the quality of the fit also the whole calibration curve can be used. The +3SD raw data signal of the zero

calibrator is calculated. Using the generated calibration curve, the LOD is equal to the concentration back calculated for the mean raw data signal of the zero calibrator plus 3SD.

- **Functional sensitivity LLOQ- ULOQ**
    - LLOQ. Lower limit of the assay that can be reported accurately based on the reproducibility of the result at a pre-specified CV. LLOQ will be established after 3-6 assay runs using results of samples generated from the standard curve calibrators spiked into most authentic matrix (plasma/serum/urine pool with low or devoid of analyte > substitute/synthetic matrix > standard solutions or buffer). Mean, SD and % CV will be calculated. Where recovery should be within 75-125%, the back calculated concentration of the lowest calibrator that does not exceed a CV 20% will be considered the LLOQ. LLOQ shall be reported as the starting concentration of the sample prior to dilution with sample buffer.
    - ULOQ: upper limit of the assay that can be reported accurately based on the reproducibility of the result at a pre-specified CV. ULOQ will be established after 3-6 assay runs using results generated from the standard curve calibrators. Mean, SD and % CV will be calculated. Where recovery should be within 75-125%, the back calculated concentration of the highest calibrator that does not exceed a CV of 20% will be considered the ULOQ. ULOQ shall be reported as the starting concentration of the sample prior to dilution with sample buffer.
- **Freeze–Thaw Stability**: Effects of multiple freeze-thaw cycles on the analyte concentration will be established using 3 different samples with low, mid and high concentrations of analyte. Samples will be divided into 4 aliquots and frozen at -70 to -80°C. After one day an aliquot of each sample (1st freeze-thaw) will be thawed at 2-8°C, mixed and returned to freezer without analysis. On the second day, the first freeze-thaw aliquot plus 1 additional aliquot sample will be thawed at 2-8°C, mixed and returned to freezer. On the third day, the first freeze-thaw aliquot, the second freeze-thaw aliquot plus one additional aliquot will be thawed at 2-8°C, mixed and analyzed in triplicate. The results of the freeze-thaw samples will be compared to the fresh sample values. Acceptable freeze/thaw stability is within the inter-assay precision for the assay and/or 25% relative error of the nominal (initial thaw) concentration.

The assessed assay validation parameters will be documented in a pre-exploratory phase validation report.

Pre-Confirmatory phase

Required experiments in the pre-confirmatory phase (for "high-bar validation"), these experiments are mainly designed following the individual CLSI guidelines for assay validation, as recommended by the regulatory agencies:

- **Calibration curve validation** cf. pre-exploratory phase
- **Recovery**: cf. pre-exploratory phase; for high bar validation recovery determination is preferably done by spiking different amounts of native human samples with high analyte concentrations into samples containing no or low amount of the analyte, if available. At least three different human samples need to be tested.
- **Freeze Thaw Stability**: cf. pre-exploratory phase
- **Linearity**

  Dilutional linearity: Use >/= 5 different individual samples of relevant sample type (plasma, serum or urine) with a measureable concentration (preferably in the higher part of the working range). Multiple samples are requested in order to consider the matrix differences that can exist between samples. The assay sample diluent is usually used for preparing the dilutions. Evaluation must include at least one native or spiked samples with a concentration above the ULOQ (to evaluate hook effect) if possible 100- to 1000-fold greater than the ULOQ. When this not feasible, efforts should be made to make the concentration as high as possible. Dilutions of the sample matrix shall be selected to cover the assay range. Samples are measured in triplicates, the recovery should be within 80 – 120 % of the expected concentration.

  Each run should contain a calibration curve, a zero standard and >/= three levels of QC-samples covering the anticipated working range (e.g. the one used to assess the intra- and inter-run precision and accuracy). Recovery should be linear for the different dilutions assessed (see "Evaluation of linearity" below). If not, a maximally allowable samples dilution must be defined in order to prevent systematic errors in analyte quantification through lack of dilutional linearity.

  Linear assay range: Linear assay range assesses the dose response relationship of the assay based on the raw signal. Ideally, the assay is tuned so that sample analyte concentrations fall within the assay linear range as this is the optimal measuring range, but it is acknowledged that often it might not be feasible. Generally, for immunoassays the linear range is usually smaller than the functional sensitivity. Linear assay range can be determined from dilutional linearity experiments if designed accordingly

  Evaluation of the linearity: Generally speaking, linearity is given when a mathematically verified straight-line relationship between raw signal and the back

calculated concentrations or the back calculated concentrations values and the true concentration of the analyte respectively, is given.

If appropriate the **assay linear range** should be determined from 7 to 11 levels of sample concentrations (measured at least in duplicates) over a range that is 20 to 30% wider than the anticipated measuring range to be able to drop off points to discover the widest possible linear range. Five points are the minimum number to reliably describe the linear range using the polynomial method described in CLSI document EP06-A. When the experiment for determination of the dilutional linearity is carefully planned, it can be combined with the assessment of the linear range of an assay. The rational and a detailed method for determination of the assay linear range are given in part B of this document.

- **Precision:** This experiment is designed to determine the total imprecision of a procedure and is based on the CLSI guidelines for the Evaluation of Precision Performance of Quantitative Measurement Methods EP05-A2. A minimum of 20 operating days is recommended, but based on assay performance, less is acceptable. On each operational day, perform two analytical runs with >/= 2 alternating operators (if it is impractical to perform two runs each day, either extend the number of operational days to n > 30 or increase the number of test samples used in each run to n ≥ 3). Use at least samples at (minimum of) two differing analyte concentrations (one should be close to the medically important decision level if applicable) and >/= 3 QC sample, all in duplicate. Routine lot to lot reagent changes as foreseen to occur during confirmatory phase must be included in precision analysis. Perform the analytical run according to the established SOP. Use routine QC procedures and materials, details are given in the SAFE-T document "*Quality control of Biomarker Immunoassay Testing*". Data may not be rejected without valid justification. If a run is rejected, conduct a replacement run only after an investigation is conducted to identify and correct the cause of the problem. State in the experimental write up if a single lot of reagents and calibration materials were used (multiple lots will provide a better estimate of total imprecision). Test materials should be as similar as possible to the usual clinical material (stable frozen pools are preferred, recombinant protein may also be used). Change the order of analysis of test and QC materials each run. Also include 10 additional samples on each run if possible to simulate actual operation. Outliers are considered as those if the difference between replicated exceeds 5.5 x the SD determined in the preliminary precision test (cf. pre-exploratory phase). If an outlier is found, the pair should be rejected and an investigation into the cause must be conducted. Calculate the repeatability SD of within run analysis using the duplicate measures of sample i. Calculate the between run SD if a "dual run" per operational day experiment is performed using data from each of the two daily runs for sample i. Calculate the day to day SD using the mean of all values for sample i from day x.

Then calculate the total (im)precision SD_{tot}. Additional statistical guidelines can be found in CLSI guidelines for the Evaluation of Precision Performance of Quantitative Measurement Methods EP05-A2. Criteria for acceptable performance may be evaluated using the degrees of freedom method (see EP05-A2) or $SD_{tot} \leq 1/3$ TEa (total allowable error).

Analytical sensitivity.. The main parameter which must be considered when establishing the lower limitation or, analytical sensitivity, is the LLoQ. Within SAFE-T, measured analyte levels below the LLoQ are usually not reported. Determination of LLoQ is described in detail in sections B.5 and B.6. In certain cases, the limit of blank (LoB), limit of detection (LoD) also need to be considered. Within SAFE-T a large number of assays need to undergo high-bar validation, thus it was agreed that LoB and LoD will only be determined if the measured analyte concentrations of well performing biomarkers are close to the lower assay analytical sensitivity. For these cases the detailed description of CLSI document EP17-A are listed as guidance in section B.12 of this document The principal experimental setup for determination of LLoQ is equal to the low bar procedure but with extending the number of requested runs to 6-10. This procedure is based on the "Guidance for Industry: Bioanalytical Method Validation" (FDA 2013). The LoB and LoD will only be determined and reported if the measured medical decision level of an analyte is close to the lower end of the analytical sensitivity of an assay. If required LoB and LoD should be determined as described in the CLSI document EP17-A. Advice for the experimental set-up following this guideline is given under B.12 of this document.

- **Lot to Lot variability:** Records should be kept of changes in assay reagent lot numbers and subsequent effects on assay day to day performance. This must be captured in the documentation used for sample testing. Routine lot to lot reagent changes as foreseen to occur during confirmatory phase must be included in precision analysis.
- **Sample Stability:** Bench-top and 4°C for 2, 4 and 24 hrs samples from ≥ 4 individuals, optional short-term refrigerator (o/n, after 2, 3 and 7 days) if required by screening laboratory procedures, short-term freezer conditions should be considered. Urine sample processing: "neat" versus supernatant.
- **Sample storage:** Establishment of long term frozen stability. Freshly collected samples and/or pools are aliquoted as soon as possible after collection and stored at -70 to -80°C and/or -20°C, following the SAFE-T SOPs for sample collection and handling. Sample with low, mid and high analyte concentrations from ≥ 5 individuals (in order to consider the matrix differences that can exist between samples) will be assayed after storage of approximately 1, 3, 6, and 12 months frozen. The percent

relative error for each storage time point will be calculated relative to baseline storage value. Acceptable long term storage stability: measured concentration is within 25% relative error of the nominal (initial thaw) concentration. Lot to lot variability needs to be considered whenever applicable for sample storage experiments.

Analytical specificity & interferences: Establish the influence of common and likely biological contaminants. >/= 5 patient samples of the relevant matrix and an intermediate analyte concentration get pooled. Preferably the experiment is performed at two analyte concentrations, one of which is close to the medical decision level and the other in the pathological range. The analyte concentration can be adjusted by "spiking" the pool with exogenous analyte. The sample pool then gets split into a test pool and a control pool. The addition of one or more potentially interfering substances to the test pool follows. The interferents are added at two different concentrations, corresponding to levels that are expected to be physiologically present, if this is feasible. The number of replicates to be taken of the test pool and the control pool is 5. Recommendations for concentrations common interferents to be tested are given in appendix B of CLSI document EP07-2A. For preparation of stock solutions of potentially interfering substances see Appendix G of the guideline. Sample size depends not only on the statistical power required to reject the null-hypothesis, but also on the maximum allowable interference to be detected at the analyte test concentration (d_{max}). This is a tool, which according to EP07-2A, has to be taken into account in order to evaluate whether an interfering effect is large enough to affect its *clinical* use. This can be evaluated through total imprecision of sample measurements, from clinical experience, or from the physiological variability. The d_{max} is hence difficult or unfeasible to assess for novel markers. It is hence suggested to base interference testing primarily on the statistical significance and to assess clinical significance at a later stage.
Advice for the analysis of results and data interpretation based on EP07-2A guideline is given under B.14 of this document

- **Estimation of total analytical error:** It is appreciated that it is not feasible in most cases to have a reference method or alternative method, given the novelty of the biomarkers being investigated in SAFE-T.

  Should a reference or alternative method be available, estimation of total analytical error should be performed: 120 patient samples are measured using the assay and a reference method. Technical replicates are not required. Results' analysis consists of a difference plot of the difference between the assay measurements and the measurements of the control method against the two methods. If the comparison method is not a reference method, the mean of the two measurement methods is used.

The width of the data band will be a rough indication of total analytical error. Data can also be analyzed using a mountain plot. The differences in the measurements of the test and the comparative method are here ranked and converted in percentiles, where percentile= rank x 100 / (N+1). To get the folded plot, the following transformation is added for all percentiles over 50: percentile= 100 – percentile. The percentiles are then plotted either as the differences or the percent differences, resulting in a frequency distribution. If the mountain plot is contained within horizontal and vertical specification lines, the total analytical error goal has been met.

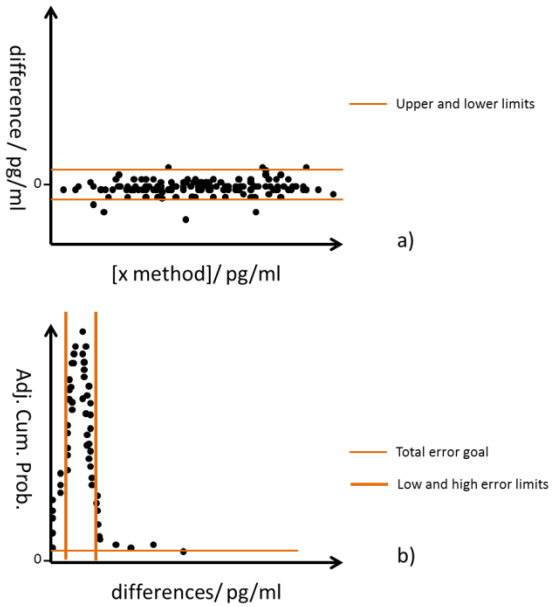

Figure 1 | a, displaying an example of a difference graph with difference plotted against the mean of the test and a reference or control method. b, displaying a mountain plot in which the percentile difference of two methods is plotted against the differences.

- **Cross-lab validation**

- **Documentation and reporting**

The documentation associated with assay entering the validation process should include:

- Application characteristics: i.e. required specimen volume, type (i.e. serum, plasma, urine etc) and specifics of sample requirements (i.e. special buffer

compositions/Protease Inhibitors/plain etc.), time of analysis, run size, equipment needed, personnel, and safety concerns.

- Methodology characteristics: i.e. test principal, reagents, optimal method for test including reaction conditions, choice of blanks, calibrators, positives/qc material, choice of reference method (if any), measurement principal and method of data reduction.
- Detailed working procedure

The assessed assay validation parameters will be documented in an exploratory phase validation report and should be published on the SAFE-T project platform.

Validation report template V2.pdf

Confirmatory phase

Required actions:

- Quality control plan, identify QC materials, establishment of a primary calibrator.
- Specify total allowable error (TEa) for the analyte at a medically important decision level (Wu, 2006).
- Finalization of test procedure and the respective SOP.
- Documentation

Possible actions:

- If for some reason it is necessary to change one assay for another one analyzing the same biomarker during the confirmatory phase, a method comparison bias estimation using patient samples should be performed, preferably according to guidelines in CLSI EP09-A2.

PART B – RATIONAL FOR VALIDATION ITEM

B.4 Assay principle

Immunoassays are based on antibodies showing affinity for a specific target (or group of targets). Quantitative determination of a biomarker of interest by immunoassays is achieved

by the quantification of the formation of the target/antibody complex formally described by the Mass Action law:

$$Ab + Ag \rightleftharpoons AbAg$$

Enzymatic immunoassays (EIA) are distinguished from other immunoassays such as Radio-immunoassays by labeling either the antigen or the antibody with an enzyme. This label in conjunction with a suitable substrate produces the assay signal allowing for the detection of the antibody-antigen complex. In the case of fluorescent immunoassays (FIA), a fluorescent component is used to detect the formation of the complex.
Nowadays the terms EIA/FIA encompass an array of different technologies or experimental setups, however all technical variations share the same general theoretical frame and assay performance can be described using the same parameters. Within the SAFE-T project the following types of immunoassays will be used by the partners:

- Enzyme linked immunoassays (ELISA): In this type of immunoassay one of the reagents is linked to a solid phase (heterogeneous), typically a 96 well plate.
- Mesoscale Discovery platform: A proprietary type of sandwich immunoassay where detection is achieved by the use of an electrochemoluminescent probe. This platform allows multiplexing of assays up to 10 analytes.
- Luminex xMAP bead-based immunoassays: A proprietary solid-phase immunoassay technology where antibodies are immobilized on colour coded beads. Detection of the Ab-Ag complex is achieved using fluorescent probes. In principle the Luminex technology allows multiplexing of assays up to 500 analytes. However, due to cross reactivity issues of the different detection antibodies combined with an increase of background signals based on the increasing concentration of the detection antibody mixture, multiplexing of sandwich immunoassays is usually limited to approximately 20 parameters.

Immunoassays like all analytical methods give a fixed response to a given amount of analyte. A characteristic of immunoassays is the inherently nonlinear relationship between response level and analyte concentration. This relation is likely to be affected by large number of parameters and may be a significant risk for the reliability of the assay.
Method validation includes all of the procedures that demonstrate that a particular method used for quantitative measurement of analytes is reliable and reproducible for the intended use. The fundamental parameters for this validation include accuracy, precision, parallelism, selectivity, sensitivity, reproducibility and stability. Validation involves documenting, through the use of specific laboratory investigations, that the performance characteristics of the method are suitable and reliable for the intended analytical applications. The acceptability of analytical data corresponds directly to the criteria used to validate the method.

B.5 Pre-exploratory phase ("low-bar validation")

During this pre-exploratory study, issues such as adequate supply of reagents, existence of a certified reference standard for the biomarker and the availability of an analyte test matrix should be fully addressed prior to taking the assay forward. Reagents, sample matrix and substitute matrix for calibrators, if appropriate, should be acquired and evaluated for their initial behavior with regard to assay selectivity, parallelism, and range of quantification with calibrator matrix.
It is recommended that different individual matrices, spiked with a known amount of reference material, should be tested during the feasibility study in recovery experiments.
Ideally neat matrix should be used but it may be difficult, particularly for multiplexed immunoassays to obtain samples without endogenous level of the analyte of interest. A first determination of the assay working range is performed. The "precision profile" (which is a plot of the coefficient of variation (CV) of the calibrator concentrations vs. the true concentration in log scale) could give an estimation of the lower and upper quantification limits. This could then be useful to check if an assay is capable of measuring the analyte of interest at some predetermined concentration range, according to the target range. If there are statistical differences in the ranges of healthy and disease populations, the method assay range must cover all the expected levels.

Samples will be tested, pure and after dilution, to assess the endogenous value of selected biomarker and to evaluate the parallelism/dilutional linearity of the assay. The minimal required dilution (MRD) may then be defined. It corresponds to the amount of dilution required to sufficiently attenuate the matrix effect, with a specific reference standard and sample diluents.

The main decision criteria to proceed forward with the validation of the assay are:

- The capacity of the assay to show a workable calibrator profile in presence of the matrix of interest.
- The capacity of the assay to cover the range of concentrations of the biomarker (taking in account potential dilution of the matrix of interest).

If a commercial reagent or kit is used, information about specificity, calibration response and sample stability, usually available in the vendor brochure should be confirmed with experiments using in-house generated samples independent of the kit reagent.

This feasibility phase and the preliminary results obtained, lead to the formulation and writing of the assay validation plan.

Additional experiments could be performed if needed.

B.6 Pre-Confirmatory phase ("high-bar validation")

During the pre-confirmatory phase, analyte ranges on different cohorts are being obtained. If those data suggest clinical usefulness of analyte levels, the assay will be assessed more thoroughly by re-addressing assay parameters from the pre-exploratory phase in more depth as well as new parameters, e.g. operator-to-operator and lot-to-lot variability. Altogether, high bar validation will need to evidence reliability of the future confirmatory data and as such, substantiate claims of clinical use. Pre-confirmatory validation procedure reflects relevant CLSI documents; if applicable In the case of an alternative approach for assessing certain validation parameters, the rational for the deviation is given.

B.7 Generation of samples to be used during the validation steps

Validation samples and Quality Control (QC) samples are samples used, respectively, during the validation process and the testing of samples to ensure that the performance and quality standards of the assay are maintained across the different runs.

Ideally validation and QC samples are clinical samples fulfilling the below criteria:

- Several (3-6) samples with known analyte levels spread over the range of interest
- Sufficient supply to perform validation and testing experiment
- Known stability of the analyte after long term storage

A reasonable alternative would be a matrix which approximately matches the typical specimen composition, such as homologous human serum-based matrix made from pooled analyte-free human serum. However, some techniques utilized for removal of endogenous analytes frequently result in major alterations in the lipid and/or protein concentration and/or composition. Affinity chromatography, in contrast, selectively removes the analyte without significantly altering the matrix. But, it is a costly procedure and is generally used only for removal of endogenous analyte present in extremely low concentrations.
Charcoal stripping, dialysis and affinity chromatography techniques are accepted as matrix preparation for the validation of assay within SAFE-T, provided that the striping method achieves a concentration of analyte which is undetectable by the method being validated. Several suppliers may provide processed plasma and serum (e.g. Innovative Research, www.innov-research.com; Seracare Life Science, www.seracare.com; Asterand, www.asterand.com).
An alternative to the use of stripped plasma is the use of animal serum/plasma exempt of cross reacting species. When using these methods it is necessary to control effective non

detection of the analyte by testing several sera from different animal species for cross-reactivity.
Another possibility acceptable in the SAFE-T project is the use of synthetic matrix. Synthetic serum (Serasub, CST Technologies Inc., www.cstti.com) and synthetic urine (Surine, CER-720, LGC Standards), free of proteins and presenting similar osmolarity which are available. This method is acceptable for the preparation of control samples since it has the advantage of not suffering from batch to batch variation.
The difficulty of selecting an appropriate matrix for the validation of the assay is exacerbated in the case of multiplexed immunoassays. Although the aforementioned strategies are still acceptable in the early stages of the project, it is strongly recommended to get in touch with regulators for use of those technologies at later stage of the project (Rodriguez et al., 2010).

B.7.1 Generation of Validation samples

Validation samples are samples used throughout the validation process. The paragraphs below define the modality of preparation of QC samples and Validation samples. We recommend the use of 6 validation samples distributed over the working range of the assay.
It is recommended to have the lowest concentration equals to 0 (which constitutes the low anchor point) and the highest concentration above the ULOQ of the assay (which constitutes the high anchor point).
Ideally validation samples are patient samples of known concentration. It is however difficult to obtain clinical sample spreading on the range of interest in sufficient quantities, therefore it is commonly accepted to generate validation sample by spiking the matrix of interest with the analyte.

B.7.2 Generation of QC samples

The ideal calibrator solution or matrix would be one whose composition is fully defined and would exactly match the composition of the unknown specimen except for the unknown analyte. These conditions are rarely attainable in practice due to the wide variability of normal and abnormal physiological specimen composition.
The preferred quality control samples are low, medium and high patient pool samples. These are favored over spiked samples, but also often unavailable. The concentration should be estimated from the data obtained during the preliminary pre-exploratory study.
QC control samples can be prepared by spiking a neat matrix at the desired level. In addition, a known endogenous level of unspiked/untreated samples can also serve as quality control.
QC sample concentrations should be assessed during the validation steps using freshly prepared samples unless it is shown that the analyte is stable over time (see section on stability). If the analyte is shown to be stable, a batch of QC control samples can be prepared, aliquoted and stored. An aliquot is thawed and assayed with each assay thereby decreasing the

possible variability associated with the preparation of the QC samples. QC samples concentration should be assessed using triplicate measurements.

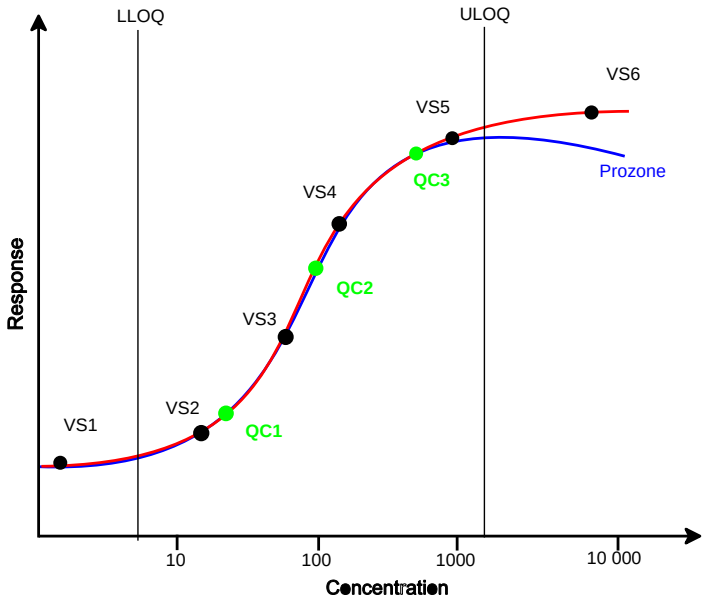

Figure 2 | This theoretical immunoassay response curve shows possible position for validation samples and QC samples. VS1 and VS6 constitute the low anchor point and the high anchor point, respectively.

B.8 Calibration curve

B.8.1 Validation of the calibration model

Most immunoassay concentration response functions are properly described by appropriate selection of the calibration model and its verification are a regulatory requirement. Therefore the first step of assay validation should be verification of the selected calibration model to ensure it adequately describes the relationship between response variable and analytical concentration, in each of the matrices studied.
The following assumptions will be made for the validation:

- The curve chosen correctly describes the data.
- The concentration data are known without errors.
- The assay errors are independent of one another.

- The variance of response values is relatively constant (homogeneity of variance or homoscedasticity).

The validation will ensure that those 4 assumptions are not grossly violated. The validation of the calibration curve should be done by running several independent experiments including the calibration curve in duplicate, during the pre-exploratory and exploratory phases. During the exploratory phase, it is required to add 2 duplicated anchor points which are standard concentrations below the lower limit of detection (low anchor point) and above the upper quantification limit (high). Those points can greatly improve reproducibility and quality of the overall curve fitting of the concentration response curve (Findlay & Dillard, 2007). Many commercial assays already feature points in the calibration curve that may be considered anchor points.
Once the data is collected, validity of the standard is checked by pooling the data of all runs and plotting the raw residual against the concentration.

$$Residual = Observed\ value - Fitted\ value$$

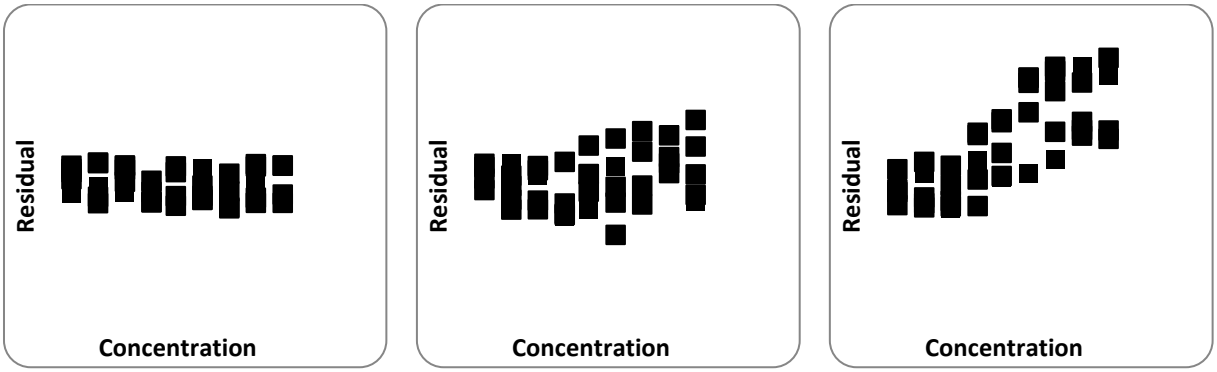

Figure 3 | Residual plots

On the figure above three plots of residual against concentration from 6 independent experiments show three typical situations. The left panel shows a theoretically ideal plot, the middle panel indicates a fitting where variance is not homogenous (the gap between points is getting larger at higher concentrations). The latter case is still acceptable but this will lead to an assay where precision will not be similar through the range of concentration. The right panel suggests a situation where the fitting is not appropriate in the higher part of the range; at this point it is appropriate to consider further assay development to find an appropriate fitting model before proceeding with the method validation.

B.8.2 Validation of the calibrator

Very often it is impractical to use the sample matrix as diluents for the calibration standard because of endogenous level of the studied biomarker. This is particularly true in the case of

multiplexed assays where stripping of all interfering endogenous factors by immune-precipitation or charcoal stripping may be impossible. For this reason it is necessary to validate the use of a synthetic buffered substitute or the use of substitute animal sera for the calibrator.
A possible method for validation in the case of a synthetic matrix is the demonstration that the biological matrix does not affect the linearity of the curve in spike recovery experiment. For the use of animal sera it has to be ensured that no cross reactivity occurs, additionally linearity of the curve has to be proven similar to the use of synthetic matrix. Further description of this methodology can be found in the section "Parallelism, Linearity".
The results of the calibration curve experiment should be reported in the validation report.

Calibrators or commercially available standards may exhibit critical batch to batch variability. Therefore it is necessary to calibrate the standards. Several strategies may be used:

- The assay of each batch of standard against a WHO primary calibrator (reference standards). The standard concentrations prepared from the primary calibrator and from the standard to be validated are run against each other. Each concentration of the standard is triplicated. A list of reference standards is available from the National Institute for Biological Standards and Controls (www.nibsc.ac.uk).
- In the absence of a WHO primary calibrator (which will be the case for the majority of SAFE-T analytes), the standard may be validated using a gold standard quantification method (e.g. HPLC). This method must be itself fully validated.
- The use of a *bridging standard*: all batches of standards are run against the same bridging calibrator and cannot cover variability between different laboratories easily.

B.9 Intra, Inter-run precision, accuracy

Definition of precision and accuracy is a regulatory requirement and provides an important benchmark for validity of generated analytical data sets as well as transfer of method.

B.10 Precision

The precision is a measure of the random error and is defined as the agreement between replicate measurements of a defined sample. It is generally expressed as the percentage coefficient of variation (%CV). Where %CV is calculated as:

$$\%CV = \frac{\sigma}{\mu}$$

with σ being the standard deviation and μ the mean.

Precision is generally expressed as intra-assay precision and inter-assay variation. The first of these is measured by comparing the analyte concentration obtained by measuring replicates of the same sample within one assay on the same day. The inter-assay variation would be determined by analyzing the same sample replicates in several different runs of the assay, with individually prepared reagents. Assay validation for the exploratory phase of the method includes demonstration of the assay performance, meeting accuracy and precision acceptance criteria within (intra-run) and between (inter-run) validation runs using the different QC samples.

QC samples are used to assess the ability of the assay to measure the biomarker of interest for its intended use, allowing one to distinguish assay variability from inherent differences amongst samples.

Operator-to-operator variability should be assessed during the pre-confirmatory phase, by assessing intra- and inter-run accuracy and precision. Operator to operator variability is assessed by evaluating the effect of the operator on inter-assay variation. It should be assessed with at least 2 operators. Inter-assay variability should be comparable from one operator to the other.

During the confirmatory phase a more robust assessment of total analytical (im)precision is made. This experiment evaluates repeatability, within day and day to day precision and then combines these to estimate the total analytical precision. Acceptable analytical performance is then determined wither by the degrees of freedom method or if SD_{tot} < 1/3 total allowable error.

Guidance for preparation of VS and QC samples is given in section B.7.2.

B.11 Accuracy

Disclaimer: for most of the SAFE-T analytes, accuracy will not be addressed due to lack of reference method or gold standard. In most of the cases, only relative accuracy can be achieved.

The accuracy is a measure of the systematic error or bias and is defined as the agreement between the measured value and the true value. Accuracy is usually reported as % bias which is calculated as:

$$\%Bias = \frac{measured\ value - true\ value}{true\ value} \times 100$$

When working with real samples, the true value is not known, making the determination of the extent of assay bias difficult.

The first line approach to determine the “true value” is the quantification of the marker using a reference method (gold standard) to quantify the marker and compare the results of the reference method to the results obtained with the assay being validated. Obviously, the reference assay has to be fully validated. This method is particularly useful when validating a multiplexed immunoassay for analytes which already have a reference single analyte assay. However, in the cases where no other assay is available for the intended application,

comparison to a reference method is not possible. One of a number of approximate methods therefore has to be used. One approach that is adopted widely is to employ the data for spiked controls generated in the assessment of the assay precision. Comparing the overall mean obtained from the repeat analysis of spiked control samples with the expected values is the normal method for calculating percentage bias.

B.12 Working range / Functional sensitivity (LLOQ-ULOQ), Limits of detection and quantification

The working range of an assay is generally defined as the highest and lowest calibrator concentrations which can be determined with an acceptable degree of precision (CV<20%).
Although data can be employed from assays containing real samples, to ensure accurate definition of the precision across the calibration range, the precision profile is best generated by repeated analysis of a calibration series (typically > 3 times). To obtain clear limits for this working range, it is recommended to include these high and low anchor points (B.8.1).
It is important however, to fully investigate and define the lower limitations of an assay (at least the LLoQ) where appropriate samples can be generated for these determinations. This allows complete understanding of the capabilities and indeed, the analytical measurement range of an assay.
The LoB is the highest apparent analyte concentration expected to be found when replicates of a blank sample containing no analyte are tested. To establish the LoB, use a minimum of 60 (NB = 60) blank measurements made up of several different blank samples i.e. 5 blank samples with 12 replicates each. This will ensure that any sample containing a noticeable level of analyte can be omitted.
Calculation of the LoB is based on the use of parametric procedures: $LoB = \mu_B + 1.645\ \sigma_B$ or $LoB = Mean + c(1—\alpha)\ SD$ where μ_B = Mean and σ_B = Standard deviation. The correction factor (1.645) is applied when estimated standard deviation (SD_s) is used in the LoB calculation as SD_s is a biased estimate. For example, for 60 (Ns = 60) measurements made with 5 (K = 5) blank samples and where f = degrees of freedom of the SD_s: $f = Ns – K$ and $c_B = 1.645/(1\text{-}1/(4 \times f))$ so, f = 60-5 = 55 and 1.645/(1-1/220) = 1.653 and therefore, $LoB = \mu_B + 1.653\ SD_B$. If the data is not normal i.e. asymmetric distribution with blank values truncated at zero (only positive results reported i.e. negative results are assigned a zero value, non-Gaussian), use nonparametric procedures: $LoB = N_B(p/100) + 0.5$ or $LoB = P(1—\alpha)$ where N_B = number of replicates and p = percentile. When a non-integer value is obtained from this equation, perform linear interpolation of the ranked data. For example, for 60 measurements in the 95th percentile (p) : $N_B(p/100) + 0.5) = 60(95/100) + 0.5 = 57.5$ so, the 95th percentile is a combination of the 57th and 58th observations and therefore, $LoB = X_{57} + 0.5(X_{58} – X_{57})$. The 95th percentile is based on the probability of false positive (Type 1 (α) error) and false negative (Type 2 (β) error) results when calculating the LoB and LoD, such that: $p = 100 – \alpha$ where $\alpha = \beta = 5\%$.

The LoD is the lowest analyte concentration likely to be reliably distinguished from the LoB and at which detection is feasible. To establish the LoD, use a minimum of 60 low level (~4 x LoB) measurements made up of several different samples i.e. 5 low-level samples with 12 replicates each. Multiple samples are recommended in order to consider the matrix differences that can exist between samples. The 60 measurements should be completed over several days so as the performance of this method can be assessed over a range of conditions, including changes in reagent lots where possible.

The LoD is calculated as: LoD=LoB+1.645(⟦SD⟧ (low concentration sample))

When LoB and LoD are determined following the EP17-A document, the LLoQ should also be determined as described in the CLSI guideline. Hereby the LLoQ is the lowest concentration at which the analyte can not only be reliably detected but at which some predefined goals for bias and imprecision are met.

Estimate the total error of the LoD as follows: Total Error = Bias + 2 x SDs where Bias = the average of the differences between the mean of the replicates and the accepted reference value of each sample and SD_s = an estimate of the pooled sample precision. To calculate the pooled sample precision: $SD_s^2 = (n_1SD_{s1}^2 + n_2SD_{s2}^2 + n_3SD_{s3}^2 + ...+n_nSD_{sn}^2) / (n_1 + n_2 + n_3....+n_n)$ where n_1, n_2, .refer to the degrees of freedom for the subsamples ($n_1 = N_1-1$, etc.). If the total error is less than the defined allowable error, then the LoQ is established and, LoQ = LoD.. This normally results if the same sample concentration is used to establish the LoD and LoQ. Otherwise, any sample concentration above the LoD which can meet predetermined goals for total error, should be used. In summary, the LoQ can be equal to or greater than the LoD. It can never be lower than the LoD.

The procedures for determination of LoB, LoD and LoQ described here (from CLSI EP17-A) are also implemented in software packages like StatisPro (http://www.statispro.org/).

B.13 Linear assay range, Parallelism & Linearity

The assay's concentration range where the raw signal is proportional to the analyte concentration is termed the linear assay range. The linear assay range is considered the most responsive and provides the most reliable quantification. Ideally the assay is tuned such that analyte concentration of biological samples falls within this range. The linear range is usually smaller than the functional assay range. It is acknowledged that the dose response curve of a large number of immunoassay reflects a sigmoidal response characteristic and not a strict linear analyte-signal response behavior and still allow precise determination of sample analyte concentrations. Regulatory authorities like to see the assessment of the assay linear range for an estimation of the most precise assay range. If the experiments for determination of dilutional linearity or parallelism are designed carefully, the linear range can be determined from experiments too.

The method to demonstrate the linearity consists of two parts. The first part examines whether a nonlinear polynomial fits the data better than a linear one. The second part, performed in cases when a nonlinear polynomial fits the data better than a linear one and assesses whether the difference between the best-fitting nonlinear and linear polynomial is less than the amount of the allowable assay bias.

For data evaluation, plot the mean value for each set of replicates at each level data on XY graph and assess visually for nonlinearity. The polynomial regression analysis for first-, second-, and third-order polynomials is performed using as many of the samples used for determination of the dilutional linearity as possible.

Order	*Polynomal*	*Regression df(Rdf)*	*nonlinear coefficient*
First	$y=b_0+b_1x$	2	
Second	$y=b_0+b_1x+b_2x$	3	b_2
Third	$y=b_0+b_1x+b_2x+b_3x$	4	b_2, b_3

The first order model is a straight line. This is the equation for the best-fitting line whether the method is linear or not. The second order model describes a relationship where there is a curved response; the third order model fits situations where the response is changing across levels ("sigmoid", s-shaped).

Obtain the standard error of the slope for each nonlinear coefficient (=SE_i, available from regression program output). Perform a t-test to test whether the nonlinear coefficients are statistically significant, that is whether the nonlinear coefficients are significantly different from zero. The test is calculated as follows, for b_2 and b_3: $t = b_i/SE_i$.

Calculate the number of degrees of freedom from the formula *df=L·R-Rdf* (L: number of different sample concentrations, R: number of replicates at each concentration, Rdf: number of degrees of freedom consumed by the regression analysis). Rdf is the number of coefficients in the regression model (including b_0).

Example: equidistant dilutions of a sample, final concentrations are 20, 40, 60, 80, 100 ng/mL, duplicate measurements, third order polynomial for fitting.

In this case: L=*5*, R=*2*, Rdf=*4* | df=*5·2-4=6*

Look up the critical value for t (two-sided at $\alpha=0.05$) in a t-table. If none of the nonlinear coefficients, b_2 or b_3, are significant ($p > 0.05$ for all), then the dataset is considered linear and the analysis is complete. If any of the nonlinear coefficients are significant ($p < 0.05$), then the dataset is considered as nonlinear. It needs to be noted that this is merely a test of statistical significance, and indicates only that nonlinearity has been detected, not that it is large enough to affect patient results.

In the case of detected nonlinearity, pick the second or third order (nonlinear) polynomial with the best fit by examining the standard error of the regression ($S_{y,x)}$.This statistic is a measure of the difference between the measured results and the model, so the model with the lowest value of $S_{y,x}$ provides the best fit for the data. Calculate the deviation of the best

second- or third order polynomial from linearity (DL_i) at each concentration as follows: $DL_i=p(x_i)-(b_0+b_1x_i)$ where the values of x range from $x_1 \ldots x_s$ and $p(x_i)$ is the value of the best-fitting polynomial at point x_i. Therefore, DL_i is the difference between the second (or third) order model and the first order model at every concentration level. This is a measure of the difference between the nonlinear model and the linear model, at each of the concentrations measured. The difference is expressed in analyte units for comparison with predefined goals. The DL_i is calculated only at the sample levels and can also be expressed as percentages.

Examine the DL_i at each level and compare with the stated criteria for error at each level. If every DL_i is less than the criterion, then although statistically significant nonlinearity has been detected, it is not important since the amount of nonlinear error is within the goal. If any DL_i exceeds the criterion, there is a possible problem with nonlinearity at that level.

Nonlinearity may be caused by sample preparation, interference, instrument calibration etc. Examine the graph response vs. concentration and determine whether the nonlinearity is at either end of the range of concentrations or in the middle of the range. If the nonlinear concentration is at either end, one option is to remove the point where DL_i was too large, and re-run the statistical analysis (see figure below linear vs. third order polynomial fit, reduction of data points considered). The linear assay range is reported as a concentration range of the analyte. The procedure for determination of assay linear range described here (from CLSI EP06-A) is also implemented in software packages like StatisPro (http://www.statispro.org/) for easier evaluation of this parameter.

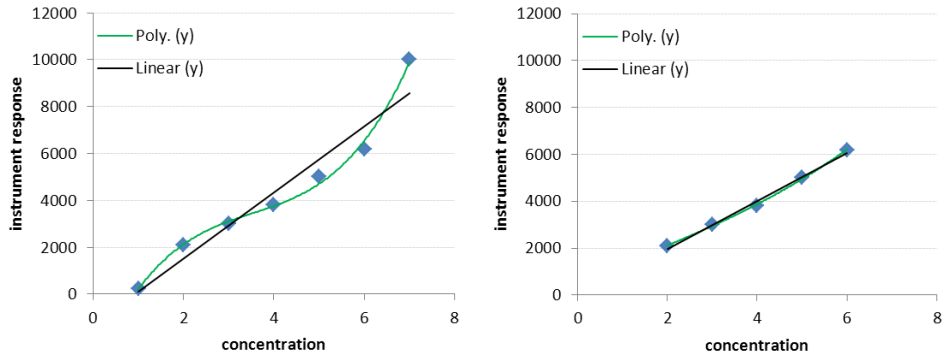

Figure 4 | Visualization plot for determination of assay linear range. Comparison of linear fit vs. third order polynomial fit. Right: Nonlinearity detected between the two models caused by samples at both ends of the curve. Left: After removal of the two outer data points, the difference between the two models for each data point is less than allowable assay bias and the linear assay range is determined by the remaining data points.

Parallelism documents the effect of the matrix on the concentration-response function of the assay. If dilution of study samples is anticipated during sample analysis, linearity of dilution must be demonstrated. It is necessary to show that the analyte of interest, when it is present in concentration above the ULOQ can be diluted to bring the analyte concentration into the

working range of the assay. An additional reason for conducting dilutional experiment is the identification of a possible “hook effect”.

B.14 Specificity/Selectivity/Interference testing

Selectivity is the extent to which the analyte may be determined without interference from other components in a mixture.

The terms selectivity and specificity are often used interchangeable. The term specific generally refers to a method that produces response for a single analyte only, while the term selective refers to a method that provides response for a number of chemical entities that may or may not be distinguished from each other. If the response is distinguished from all other responses, the method is said to be selective.

Validation of selectivity is necessary to confirm that the known or presumed presence of different substances in a mixture does not interfere with the measured analyte value.

Selectivity may be determined by measuring the analyte in absence and presence of other compounds (aqueous sample versus matrix). This is especially important as the screening of patient samples includes a diverse population. A newly developed method should not be considered suitable for the analysis of samples before matrix interference experiments have been conducted. In particular, the presence of compounds like haemoglobin, bilirubin, fat emulsion, rheumatoid factors, triglycerides, HAMAs (human anti mouse antibodies), as well as other matrix effects of human samples which may interfere with immunoassays has to be considered during assay development and validation. This may lead to a need for diluting samples prior to analysis and/or to use specific buffer systems including appropriate blocking reagents (e.g. animal sera, heterophilic antibody blockers, etc.).

The selectivity of an assay in the presence of endogenous substances should be evaluated by using neat matrices from multiple individuals and both genders, if possible. In addition, a known concentration of the biomarker should be spiked into the various individual lots of neat matrix to assess the recovery and to test for specific matrix effects.

Cross-reactivity of endogenous compounds should be evaluated individually and in combination with the analyte of interest. When possible the newly developed immunoassays should be compared with validated reference methods (such as LC-MS) using incurred samples. The dilutional linearity to the reference standard should be assessed using study samples. Nonspecific binding should be determined.

<u>Analysis of results from interference testing</u>

The observed interference effect d_{obs} is calculated as the difference between the means of the test and control samples: Interference = $d_{obs} = \bar{x}_{test} - \bar{x}_{control}$. The cut-off level d_c is computed to determine whether to accept the null-hypothesis or to reject it. The d_c value can be calculated from: $d_c = d_{null} + sz_{1-\alpha/2} / \sqrt{n}$, where d_{null} is the value stated in the null-hypothesis, usually 0. For a one sided-test, $1-\alpha/2$ is replaced by $1-\alpha$.

Interpretation of data from interference testing:

The values of d_c and d_{obs} are then compared to judge whether the potential interferent is in fact interfering or not. If d_{obs} is higher than the cut-off value d_c, there is interference. This takes only in account the statistical significance. If the experiment was designed to take into account clinical significance, then, if the point estimate, d_{obs}, is less than or equal to the cut-off value, d_c, it can be concluded that the bias caused by the potential interferent is less than d_{max}. Otherwise, the alternative hypothesis needs to be accepted, that there is interference.

If interference exists, it is characterized in terms of dose-dependency. This is done by volumetric admixtures of the sample containing the highest interferent concentration and lower concentration pools. About 5 concentrations of interferent with a constant amount of analyte are generated in this way and are measured in triplicate. The observed effect is plotted against the interferent concentration and the dose-response relationship examined. Alternative to adding potential interferent to a test sample and comparing it to control, a comparative approach can be chosen. In this, the assay is compared with a method that measured the same analyte, but is unaffected by the potential interferent (which would require a well-characterized gold standard). It is recognized that this approach will not be feasible in most cases. Interfering substances could include hemoglobin, bilirubin, serum or plasma (urine analytes only), fat emulsion (Liposyn, Abbott Labs or Intralipid, Cutter labs) (serum & plasma analytes only), myoglobin, urine preservatives (e.g. sodium azide), common prescription drugs, abnormal biochemical metabolites expected in the patient population, substances reported to interfere with similar methods, sample additive such as anti-coagulants, substances that may come in contact with the specimen during collection and processing and certain dietary substances. The interference of certain substances listed above can be ruled out on a scientific basis. A rational should be presented on why these can be ruled out. For specific diseases, the robustness of the assay with respect to rheumatoid factor (Rf) and human anti-mouse antibodies (HAMA) will need to be assessed.

B.14.1 Sample size calculations for interference testing taking into account the clinical significance

As mentioned in **A.3 Pre-confirmatory phase** under the subsection **Interference Testing**, the evaluation of interference requires statistical as well as clinical criteria. For novel markers, it is unfeasible or very difficult to take into account clinical significance. However, in cases where this is possible, d_{max}, the maximum allowable interference to be detected at the analyte test concentration, can be evaluated. If clinical significance is taken into account, the number of technical replicates has to be adjusted accordingly. Below is described a method to do so.

For a two-sided test (i.e. one in which the interference is not known to decrease or enhance the signal), the following formula is used to calculate the required sample size: $n = 2[(z_{1-\alpha/2} + z_{1-\beta})s/d_{max}]^2$.

Where

$z_{1-\alpha/2}$ is the percentile from the standardized normal distribution corresponding to the confidence level 100(1-α)% for a two-sided test.

$z_{1-\beta}$ is the percentile from the standardized normal distribution corresponding to the power of 100(1-β)%

s is the repeatability standard deviation of the measurement procedure. It is evaluated during the low bar validation as intra-run precision.

d_{max} is the maximum allowable interference to be detected at the analyte concentration. The units are the same as the ones used to quantify the analyte in the assay. It is the maximum observed sample value that can be judged to be without clinical significance, when compared to the control with no added potential interferent. (It could be for example the observed sample mean +/- 3 SD). It needs to be defined for each method.

For a one-sided test, the term $z_{1-\alpha/2}$ is replaced by $z_{1-\alpha}$ in the above equation.

The z-values for some commonly used confidence and power levels are shown below:

Confidence (Power)	0.9	0.95	0.975	0.99	0.995
z-percentile	1.282	1.645	1.960	2.326	2.576

- here $z_{1-\alpha/2}$ is the percentile from the standardized normal distribution corresponding to the confidence level 100(1-α)% for a two-sided test, where $z_{1-\beta}$ is the percentile from the standardized normal distribution corresponding to the power of 100(1-β)%, where s is the repeatability standard deviation of the measurement procedure and where d_{max} is the maximum allowable interference to be detected at the analyte concentration. A practical way of assessing the number of replicates needed is to express the imprecision first as a multiple of the repeatability standard deviation (d_{max}/s). Then, Table 1 can be used to infer the number of replicates in order to detect interference effects with 95% confidence and power.

Hence, the number of replicates chosen reflects the confidence level and power with which the null-hypothesis is tested, the repeatability of the measurement procedure and the magnitude of the smallest difference between the analyte test results that is considered clinically significant. This "clinical significance" changes with the method. It should be determined prior to testing the interference for a particular method.

Table 1: Number of replicates ruling out interference at a maximum allowable level (expressed as a multiple of the repeatability standard deviation (d_{max}/s)) with a 95 % confidence limit

d_{max}/s	No. of replicates	**d_{max}/s**	No. of replicates
0.8	41	1.5	12
1.0	26	1.6	10
1.1	22	1.8	8
1.2	18	2.0	7
1.3	16	2.5	5
1.4	14	3.0	3

Example of interference experiment (copied from CLSI document EP07-A2):

A recent kidney recipient showed a repeatable change of creatinine level from 1.0 to 1.2 mg/dL. The physician wants to know if the change could be caused by a cephalosporin antibiotic.

At 1 mg/dL creatinine, the repeatability standard deviation is 0.075 mg/dL. The laboratory considers 0.1 mg/dL a significant interference. With adequate replication, the effect of imprecision can be reduced so that a possible interference of 0.1mg/dL would be detected.

First, express the imprecision as a multiple of the repeatability standard deviation (d_{max}/s): 0.1 mg/dL / 0.075 mg/dL = 1.33 .

Then, rounding down to 1.3, the above table can be used to determine the required number of replicates. It shows that detecting an effect of this magnitude with 95% confidence and power requires 16 replicates each for the control and test conditions.

If a larger interference is considered acceptable, such as an effect of 0.2 mg/dL (dmax/s=2.7), fewer replicates would be needed to achieve the same degree of confidence. The table would then suggest that only 4 replicates are needed instead of 16.

B.15 Samples Stability

It is necessary to determine how stable an analyte is with respect to time and storage conditions. As a prerequisite a standardized sample collection procedure has to be installed. This is ensured by the SAFE-T SOPs for sample collection and handling.

B.15.1 Freeze and Thaw stability

Generally repeated freeze/thaw cycles should be avoided. If possible, aliquots of samples should be thawed only if directly used for measurements. For pre-exploratory phase, the stability will be tested after going through multiple freeze/thaw cycles, Analyte stability will be determined after three to five freeze and thaw cycles.
Native samples or a substitute matrix spiked with 3 different analyte concentrations (low, mid, high): low (target concentration close to twice the LLOQ), mid and high (half of the ULOQ) will be aliquoted into 3 vials and frozen immediately. The samples should be stored for at least 24 hours at the intended storage temperature. Then, one of the aliquots of each sample preparation should be thawed unassisted at room temperature while the other two should be kept frozen as controls. When completely thawed, the samples should be refrozen for 12 to 24 hours under the same conditions. The freeze-thaw cycle should be repeated two to four more times. Then all samples will be thawed, analyzed and values obtained will be compared to those of the controls which did not undergo the repeated freeze-thaw cycles.

B.15.2 Analyte and Sample storage stability

For the exploratory phase, the stability testing will evaluate the stability of the analyte after long-term (frozen at the intended storage temperature) and short-term (bench-top, room temperature) storage.
Generally stability testing can be a complex issue due to the difficulties in defining biomarker stability under storage conditions and in judging the adequacy of the assay method to monitor stability changes. Conditions used in stability experiments should reflect situations likely to be encountered during actual sample handling and analysis. The procedure needs also to include an evaluation of analyte stability in stock solution.
One should be aware that, unlike small molecules, stability measurements of protein biomarkers can be method-dependent; this implies that stability data should be interpreted with the understanding of the method in relevance to the biology of the biomarker.

B.15.3 Short-Term Temperature Stability

Samples should be thawed on ice and kept at room temperature for 2, 4 and 24 hours (Bench-top) to simulate the time samples will be maintained at this temperature for analysis. In parallel the same samples should be thawed on ice and kept at 2-8°C for 2, 4 and 24 hours to simulate short-term sample storage at 2-8°C while sample or assay preparation. After that time the baseline aliquot of the samples should be thawed on ice and all samples together should be analyzed immediately by the intended method. Additionally samples stored in a refrigerator for 2, 3, and 7 days can equally be tested if these conditions are required by the screening laboratory. Short term freezer conditions should also be considered.

B.15.4 Long-Term Stability

Ideally, the storage time in a long-term stability evaluation should exceed the time between the date of the first sample collection and the date of the last sample analysis. During the duration of the Safe-T project this will not be possible to determine. Long-term stability should be determined by storing at least three aliquots of each spiked samples (low, mid and high concentrations) under the same conditions as the study samples. The volume of samples should be sufficient for analysis on three separate occasions. Within Safe-T, samples will be stored for 1 month, 3 months, 6 months and 12 months and directly analyzed using identical analytical methods. The concentration of all stability samples should be compared to the measured concentration at the first day of storage and have a relative error < 25%.

B.15.5 Stock Solution Stability/Dilutions of Standard Proteins

The stability of stock solutions of the standard proteins should be evaluated at room temperature. Standard dilution series are prepared in the preferred matrix and kept at room temperature for at least 6 hours. After completion of the desired storage time, the stability should be tested by comparing the instrument response with that of a freshly prepared standard dilution series.

B.16 Lot to lot variability

Lots are defined as an ensemble of assays for which critical assay components are coming from the same batch (antibodies, standard reagent, diluents (if complex diluents are used). Inter-assay precision between lots should be assessed using QC samples.

B.17 pH effect

Changes in the pH-value of samples may change analyte characteristics and/or immunoassay performance. Different pH-values will not be extensively evaluated during assay

development. But it has to be ensured that the assay conditions with respect to the pH value are maintained for assay development and sample screening. In this context it is important to prepare standard dilution series under conditions as similar as possible (pH, matrix) to the samples to avoid varying signals caused by different analytical conditions.

REFERENCES

European Medicine Agency. (2009). Draft guideline on validation of bioanalytical methods. London, UK: COMMITTEE FOR MEDICINAL PRODUCTS FOR HUMAN USE.

Clinical and Laboratory Standards Institute (2003). Evaluation of the Linearity of Quantitative Measurment Procedures : A Statistical Approach ; Approved Guidelines. CLSI document EP06-A. *Volume 23* (16). CLSI, Wayne, PA, USA. ISBN 1-56238-498-8

Clinical and Laboratory Standards Institute (2004). Protocols for Determination of Limits of Detection and Limits of Quantitation, Approved Guidelines. CLSI document EP17-A. *Volume 24* (34). Wayne, PA, USA. ISBN 1-56238-551-8

Clinical and Laboratory Standards Institute (2004). Evaluation of Precision Performance of Quantitative Measurement Methods; Approved Guideline. CLSI document EP05-A2. *Volume 24* (25). CLSI, Wayne, PA, USA. ISBN 1-56238-542-9

Clinical and Laboratory Standards Institute (2003).Estimation of Total Analytical Error for Clinical Laboratory Methods; Approved Guideline. CLSI document EP21-A. *Volume 23* (20). CLSI, Wayne, PA, USA. ISBN 1-56238-502-X

Clinical and Laboratory Standards Institute (2005). Interference Testing in Clinical Chemistry ; Approved Guideline-Second Edition. CLSI document EP07-A2. *Volume 25* (27). CLSI, Wayne, PA, USA. ISBN 1-56238-584-4

Findlay, J. W. A., & Dillard, R. F. (2007). Appropriate calibration curve fitting in ligand binding assays. *The AAPS journal*, *9*(2), E260-7. doi: 10.1208/aapsj0902029.

Food and Drug Administration (2013). Guidance for Industry: Bioanalytical Method Validation. Revision 1. Rockville, MD: US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research.

Lee, J. W., Devanarayan, V., Barrett, Y. C., Weiner, R., Allinson, J., Fountain, S., et al. (2006). Fit-for-Purpose Method Development and Validation for Successful Biomarker Measurement. *Pharmaceutical research*, *23*(2), 312-328. doi: 10.1007/s11095-005-9045-3.

Lee, J. W. & Hall, M. (2009). Method validation of protein biomarkers in support of drug development or clinical diagnosis / prognosis. *Journal of chromatography B*, *877*, 1259-1271. doi: 10.1016/j.jchromb.2008.11.022.

Rodriguez, H., Tezak, Z., Mesri, M., Carr, S. A., Liebler, D. C., Fisher, S. J., et al. (2010). Analytical validation of protein-based multiplex assays: a workshop report by the NCI-FDA interagency oncology task force on molecular diagnostics. *Clinical chemistry*, *56*(2), 237-43. doi: 10.1373/clinchem.2009.136416.

Wu, A. (2006). Analytical issues for clinical use of cardiac troponin. In D. A. Morrow (Ed.), *Cardiovascular Biomarkers: Pathophysiology and Disease Management* (p. 27). Totowa, NJ: Humana Press Inc.