nature portfolio

https://doi.org/10.1038/s41596-022-00801-8

Supplementary information

A proposed framework to evaluate the quality and reliability of targeted metabolomics assays from the UK Consortium on Metabolic Phenotyping (MAP/UK)

In the format provided by the authors and unedited

Supplementary Table 1. Comparison of the U.S. Food and Drug Administration (FDA), European Medicines Agency (EMA) and International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) M10 guidelines.

Validation parameters	FDA (2018)	EMA (2011)	ICH M10/ EMA (2019)
Accuracy/trueness	• Within-run: in a single run, a minimum of 5 QCs per level at a minimum of 4 concentration levels-LLOQ, low, medium, and high. • Between-run: 3 runs of 5 QCs per level at a minimum of 4 concentration levels-LLOQ, low, medium, and high at 2 different days should be evaluated. • Mean should be within 15% of the nominal value (except at LLOQ, no more than 20%).	• Mainly the same as FDA. • Accuracy is defined by the formula: (determined value/true value) x100%.	• Mainly the same for FDA and EMA.
Precision	• Within-run: in a single run, a minimum of 5 QCs per level at a minimum of 4 concentration levels-LLOQ, low, medium, and high. • Between-run: 3 runs of 5 QCs per level at a minimum of 4 concentration levels-LLOQ, low, medium, and high at 2 different days should be evaluated. • Should not exceed 15% of the coefficient of variation (CV% or RSD%) except for the LLOQ, where it should not exceed 20% CV.	• Precision measured using 5 samples at 3 concentrations in a single run (intra-day), and 3 runs for 3 concentrations on at least 2 different days (inter-day). • Acceptance criteria is mainly the same as FDA.	• Mainly the same as FDA.

Validation parameters	FDA (2018)	EMA (2011)	ICH M10/ EMA (2019)
Lower limit of quantification (LLOQ)	• Lowest non-zero standard on the calibration curve defines the sensitivity (LLOQ). • Analyte response at the LLOQ should be ≥5 times the analyte response of the zero calibrator. • Establish LLOQ using at least 5 samples independent of standards and determine %CV or confidence interval. • Should be ≥5 times the response compared to blank response. LLOQ analyte peak response should be identifiable, discrete and reproducible with a precision of 20% and accuracy of 80-120%. • Accuracy should be ± 20% of nominal concentration (from ≥5 replicates in at least 3 runs). • Precision should be ±20% CV (from ≥5 replicates in at least 3 runs).	• Lowest concentration of analyte in a sample which can be quantified reliably with an acceptable accuracy and precision. • LLOQ is considered as the lowest calibration standard. • S/N≥5:1.	• Term has been used frequently but has not been defined. Only in the glossary has the term been described without further details of procedure or acceptance criteria.
Calibration curve standards (linearity)	• Calibrators (calibration standards): a biological matrix to which a known amount of analyte has been added. Calibration standards are used to construct calibration curves from which concentrations of analytes in QC samples and in-study samples are determined. • Should consist of a blank sample, a zero sample and at least 6 non-zero samples covering the expected range, including LLOQ. • Non-zero calibrator is a calibrator to which the internal standard is added. • Zero calibrator is a blank sample to which the internal standard is added. • No recommendation for inclusion of internal standard (IS) for MS analysis. • No recommendation for number of curves. • No recommendation for acceptance criteria for number of replicates. • Estimation of concentration of unknown sample by extrapolation of standard curve below LLOQ or above ULOQ is not allowed.	• Generate for each analyte in the sample. • Prepare in the same biological matrix as sample or surrogate matrices. • Range coverage of the expected sample concentrations (LLOQ to ULOQ, minimum of 6 levels), in replicate including blank sample (matrix sample processed without analyte or internal standard), and a zero sample (matrix sample processed without analyte but with internal standard). • Minimum 3 calibration curves for validation report (CC) and 6 for ligand binding assays (LBAs). • Stable isotope-labelled approach for MS analysis is recommended. • Acceptance criteria for replicates should fulfil ≥50% of standard per level.	• Mainly the same for FDA and EMA. • Concentration of the calibration standards should be within ±15% of the nominal value, except for the LLOQ which should be within ±20%. • At least 75% of the calibration standards should fulfil above criteria. • Use of previously prepared and stored stable standards are allowed. • In the case that replicates are used, criteria (within ±15% or ±20% for LLOQ) should also be fulfilled for at least 50% of the calibration standards tested per concentration level.

Validation parameters	FDA (2018)	EMA (2011)	ICH M10/ EMA (2019)
Internal standard (IS) and reference standard	• FDA recommends the characterisation of internal standard (IS) and reference standard (CoA; certificate of analysis). If IS or reference standard has expired, the stock solution made from the expired lot should not be used unless purity and stability is re-established.	• EMA does not require CoA for IS as long as its suitability is demonstrated.	• Mainly the same as FDA.
Selectivity/specificity	• 6 individual sources of the appropriate blank matrix tested for interference. • Absence of interfering compound accepted where the response is less than 20% of LLOQ and/or less than 5% for IS. • Back-conversion of metabolite into parent analyte has not been mentioned.	• Back-conversion of metabolite into parent analyte should be evaluated for extraction procedures and MS assay.	• Mainly the same for FDA and EMA.
Recovery	• Extracted samples at low, medium and high QC concentrations versus extracts of blanks spiked with the analyte post extraction (at low, medium and high).	• Not mentioned.	• Extraction efficiency. • Recovery of the analyte need not be 100%. • Three concentrations (low, medium and high) compared with unextracted standards that represent 100% recovery or post-spiked extracted blanks.
Stability	• Stability in auto-sampler, bench-top, extract, freeze-thaw, stock solution and long-term. • Perform at least 3 replicates at low and high QC concentrations. • The accuracy (% nominal) at each level should be ±15%. • Number of freeze/thaw cycles should be at least 3 times. • Long-term stability at -20◦C should cover stability at lower temperatures. • Impact of co-medication should be considered. • Stability of stock and working solutions should be compared with freshly prepared solutions (85-115% accuracy).	• Freeze and thaw stability (cycles should be equal to or greater than the freeze/thaw cycles intended for the study samples). • Bench-top (short term) stability. • Long-term stability (small molecules bracketing approach is acceptable; large molecules-stability at each storage temperature required). • Stock and working solution stability. • Processed sample stability/auto-sampler stability. • Low and high QC concentrations, replicates have not been mentioned. • The accuracy (% nominal) at each level should be ±15%. • Impact of other analytes should be considered.	• Stability of stock and working solutions. • Freeze and thaw matrix stability. • Bench top (short-term) matrix stability. • Processed sample stability. • Long-term matrix stability. • Whole blood stability. • Low and high QCs, at least 3 replicates per concentration level/storage condition/time point. • Mean concentration at each QC level (analytes and IS) should be within ±15% of the nominal concentration. • Ratio of means or confidence intervals in both datasets (reference and after storage) should be used.

Validation parameters	FDA (2018)	EMA (2011)	ICH M10/ EMA (2019)
Carryover	• No detailed procedure for assessing carryover. • The impact of carryover on the accuracy of study sample concentrations should be assessed. • Should not be greater than 20% of the LLOQ and 5% for the IS.	• Should be assessed by injecting blank samples after a high concentration sample or highest calibration standard. • Should not be greater than 20% of the LLOQ and 5% for the IS. • If unavoidable, study samples should not be randomised.	• Mainly the same for FDA and EMA.
Dilution integrity	• QCs for planned dilutions, 5 replicates per dilution factor. • Accuracy: ±15% of nominal concentrations. • Precision: ±15% CV.	• Spike blank matrix to concentration above ULOQ and dilute down with blank matrix (5 determinations per dilution). • Accuracy and precision should be within ±15%.	• Mainly the same for FDA and EMA.
Matrix effect	• Analytes should be quantified free of potential interfering substances including endogenous matrix components, metabolites, anticipated concomitant medications, etc. • Blank samples of the appropriate biological matrix from at least 6 individual sources should be assessed for matrix effect on ion suppression, ion enhancement and/or extraction efficiency. • Blank and zero calibrators should be free of interference at the retention times of the analyte(s) and the IS. • Spiked samples should be ±20% of the LLOQ. • The IS response in the blank should not exceed 5% of the average IS responses of the calibrators and QCs.	• Matrix factor (MF) should be calculated using 6 different sources of blank matrix. • Determination should be done by spiking to low (3 times LLOQ) and high (close to ULOQ) levels of concentration. • Should not exceed 15% CV (RSD%).	• Matrix effect should be evaluated by analysing at least 3 replicates of low and high QCs, each prepared using matrix from at least 6 different sources/lots. • Accuracy should be within ±15% of the nominal concentration, and precision (%CV) should not be greater than 15% in all individual matrix sources/lots. • Use of fewer sources/lots may be acceptable in the case of rare matrices.

Validation parameters	FDA (2018)	EMA (2011)	ICH M10/ EMA (2019)
Quality control (QC)	• For accuracy and precision runs: 4 concentrations (LLOQ, low, medium, high) in 5 replicates in at least 3 runs. • For other validation runs: 3 concentrations (low, medium and high) in duplicates. • FDA suggests freshly prepared QC samples (for at least one analysis). • Evaluate QC samples drift and its impact on accuracy. • Include QC outliers in the assessment of accuracy and precision from accepted runs.	• At least 3 levels of QC samples (low, medium and high) in duplicate (or at least 5% of the number of study samples, whichever is higher). • EMA advises frozen QCs-treated in the same way as the study samples. • EMA has no recommendations for QC drift and outliers.	• 4 QC concentrations (LLOQ, low, medium and high), but no mention on number of replicates. • Calibration standards and the QCs should be prepared from separate stock solutions in order to avoid biased estimations which are not related to the analytical performance of the method. • Calibration standards and the QCs may be prepared from the same stock solution if the accuracy and stability of the stock solution have been verified. • Endogenous concentrations of the analyte in the biological matrix should be evaluated prior to QC preparation (e.g., by replicate analysis). • Blank matrices with minimum level of endogenous analyte should be used. • Concentrations of QCs should account for endogenous concentrations in biological matrix (i.e. additive) and be representative of expected study concentrations. • QCs used for validation should be aliquots of authentic biological matrix unspiked and spiked with known amounts of authentic analyte. • In spiked samples, the added amount should be enough to provide concentrations that are statistically different from the endogenous concentration.
Reproducibility	• Reproducibility is the precision between 2 laboratories. It also represents the precision of the method under the same operating conditions over a short period of time.	• Not mentioned.	• According to the ICH guidelines, precision may be considered at 3 levels: repeatability, intermediate precision and reproducibility. • Repeatability expresses the precision obtained under the same operating conditions over a short interval of time. Repeatability can also be termed intra-assay precision. It is likely that the assay would be repeated by the same person using a single instrument. • Intermediate precision expresses within-laboratory variation of precision when the

Validation parameters	FDA (2018)	EMA (2011)	ICH M10/ EMA (2019)
			analysis is carried out by a different analyst on different days and with different equipment. Analysts should be trained to the same standard of procedure, data handling and standardised use on particular items of equipment. • Reproducibility expresses precision between laboratories.
System suitability	• System suitability is a determination of instrument performance (sensitivity and chromatographic retention) by analysing a set of reference standards prepared at a known concentration in mobile phase before the analytical run, with 3 at the beginning and 3 at the end of the run.	• Not mentioned.	• System suitability, including apparatus conditioning and instrument performance, should be determined using samples that are independent of the calibration standards and QCs for the run. Subject samples should not be used for system suitability. No further description of procedure or acceptance criteria.
Dried matrix methods (DMM)	• A microsampling technique that provides advantages of reduced blood sample, ease of collection, storage and transportation. • Should be assessed by multiple punches of the sample or samples should be taken in duplicate. • Incurred sample re-analysis (ISR) should be performed. • Cross-validation is required if other analytical techniques are performed on the samples.	• Not mentioned.	• Mainly the same as FDA.
Internal standard (IS) response variability	• Should be monitored.	• Study samples should be re-analysed if IS variability is significant.	• The IS responses of the study samples should be monitored to determine whether there is systemic IS variability.

Validation parameters	FDA (2018)	EMA (2011)	ICH M10/ EMA (2019)
Incurred sample re-analysis (ISR)	• ISR is the reanalysis of study samples in separate runs on different days to evaluate the accuracy of incurred samples. • ISR confirms the reproducibility and reliability of a validated bioanalytical method. • ISR should be 7% of the study sample size. • Acceptance criteria: two-thirds (67%) of the repeated sample results should be within 20% for small molecules and 30% for large molecules.	• Reanalysis of study samples in separate runs at different days; 10% of the samples to be reanalysed if sample numbers are less than 1000, and 5% of the numbers if more than 1000 samples. • Acceptance criteria: Percent difference between original value and repeat analysis should not be greater than 20% of the mean value for at least 67% of the repeats.	• ICH and EMA have listed the sample types that ISR should be performed on. • Acceptance criteria: if the total number of study samples is less than 1000, then 10% of the samples should be reanalysed; if the total number of samples is greater than 1000, then 10% of the first 1000 samples (100) plus 5% of the number of samples that exceed 1000 samples should be assessed. • A subset of samples should be chosen randomly, but for PK studies, the samples for ISR should be chosen around the maximum concentration (Cmax) and some in the elimination phase. • Samples should not be pooled, as pooling may limit anomalous findings. • ISR should be performed within the stability window of the analyte, but not on the same day as the original analysis. • Acceptance criteria: two-thirds (67%) of the repeated sample results should be within 20% for chromatographic methods and 30% for ligand binding assays.
Parallelism-(ligand binding assay, LBA)	• Investigate parallelism (for endogenous products), to ensure that the serially diluted incurred sample response curve is parallel to the calibration curve. • Parallelism is a performance characteristic that can detect potential matrix effects and interactions between critical reagents in an assay. • Procedure has not been described in detail.	• Parallelism between the calibration standard curve and serially diluted study samples should be assessed to detect possible matrix effect or differing affinities for metabolites. • A high concentration study sample (preferably close to Cmax) should be diluted to at least 3 concentrations with blank matrix. The precision between samples in a dilution series should not exceed 30%.	• Parallelism is defined as a parallel relationship between the calibration curve and serially diluted study samples to detect any influence of dilution on analyte measurement. • Should be evaluated in the surrogate matrix and surrogate analyte by standard addition approach, spike recovery or dilutional linearity. • A high concentration study sample (preferably close to Cmax) should be diluted to at least 3 concentrations with blank matrix. Precision between samples in a dilution series should not exceed 30%.

Validation parameters	FDA (2018)	EMA (2011)	ICH M10/ EMA (2019)
Selectivity/specificity-(ligand binding assay, LBA)	Selectivity-LBA: • Conduct an analysis of blank samples in the matrix from ≥10 individual sources. • Acceptance criteria: for ≥80% of sources, un-spiked matrix should be below quantification levels, and spiked samples should be ±25% at LLOQ, and ±20% at high QC. Specificity-LBA: • Assessed for interference by cross-reacting molecules, concomitant medications, bio-transformed species, etc. • Potential interfering materials should be added to calibration curves in buffer. • Acceptance criteria: QCs should meet ±20%, or 25% at the LLOQ and ULOQ.	Selectivity-LBA: • Ability to quantify the analyte of interest in the presence of unrelated compound in the matrix. • Test by spiking at least 10 sources including lipemic and haemolysed sample matrix at or near LLOQ. • If interference is concentration related, it is essential to determine the minimum concentration where interference occurs. • Accuracy should be within 20% (25% at LLOQ) of the nominal spiked concentration in at least 80% of the matrices evaluated. Specificity-LBA: • Specificity of binding reagent refers to its ability to bind solely to the analyte of interest, not metabolites of the analyte. • Test with QC samples by adding increasing concentrations of available related molecules or drugs expected to be concomitantly administered, and measuring the accuracy of the macromolecule of interest at both LLOQ and ULOQ. • Assay acceptance criteria: QC samples should be within 25% of the nominal values.	Selectivity-LBA: • Should be evaluated at low and higher analyte concentrations. • Obtain blank samples from 10 individual sources and spike them to LLOQ and high QC levels. • Response of blank samples should be below the LLOQ for at least 80% of individual sources. • Accuracy should be within ±25% at the LLOQ and within ±20% at the high QC level of the nominal concentration in at least 80% of the individual sources evaluated. • Evaluate selectivity for lipaemic and haemolysed samples using a single source of matrix. • In relevant patient population studies, selectivity should be performed for at least 5 individual patients. Specificity-LBA: • Specificity should be evaluated by spiking blank matrix samples with related molecules (structurally related molecules) at the maximal concentration(s). • Investigate the accuracy of target analyte at the LLOQ and at the ULOQ. • Response of blank samples spiked with related molecules should be below the LLOQ. • Accuracy should be within ±25% of the nominal values. • If the method is non-specific, then minimum concentration of the interfering molecule should be determined by spiking increasing concentrations of interfering molecules in blank matrix and measuring the accuracy of the target analyte at the LLOQ and ULOQ. • Employ appropriate mitigations during sample analysis; adjust LLOQ/ULOQ accordingly and/or consider a new method.

Validation parameters	FDA (2018)	EMA (2011)	ICH M10/ EMA (2019)
Prozone (hook) effect-(ligand binding assay, LBA)	• Hook effect: when increasing analyte concentrations results in no change or decreased signals compared to the preceding concentration.	• Signal suppression caused by high concentration of analyte.	• Dilution linearity and hook effect should be investigated. • QCs for dilution should be prepared with the same matrix as the study sample. • Spike the matrix with analyte concentration above ULOQ (undiluted) for hook effect and dilute this sample (to at least 3 different dilution factors) with blank matrix to a concentration within the calibration range. • At least 3 runs for each dilution factor- the same number of replicates that were used for the study samples. • Calculated concentration for each dilution should be within ±20% of nominal concentration after correction for dilution, and precision of the final concentrations across all dilutions should not exceed 20%. • Dilution factor(s) applied during study sample analysis should be within the range of dilution factors evaluated during validation.
Minimum required dilution (MRD)-(ligand binding assay, LBA)	• Not mentioned.	• Definition provided. • Correction for dilution mentioned.	• MRD is the dilution factor used for dilution of study sample with buffer solution in order to reduce background signal or matrix effect. • MRD should be consistent for all samples including calibration standards and QCs. • Matrix can contain non-specific matrix component such as degrading enzymes, heterophilic antibodies or rheumatoid factor which may interfere with the analyte of interest. • Partial validation is required if MRD is changed.

Supplementary Table 2. Definition, methodology, and acceptance criteria for validation parameters: lower limit of quantification (LLOQ), matrix effect, carryover, parallelism, dilutional linearity/integrity, prozone (hook) effect.

Validation parameter	Definition	Methodology	Acceptance criteria
Lower limit of quantification (LLOQ)	• Lowest concentration of analyte in a sample which can be quantified reliably with an acceptable accuracy and precision. • LLOQ is considered the lowest calibration standard. • S/N≥5:1.	• Establish LLOQ using at least 5 samples independent of standards and determine %CV or confidence interval.	• Analyte response at the LLOQ should be ≥ 5 times the analyte response of the zero calibrator. • Accuracy should be ±20% of nominal concentration (from ≥ 5 replicates in at least 3 runs). • Precision should be ±20% CV (from ≥5 replicates in at least 3 runs).
Matrix effect (Matrix factor)	• Matrix effect refers to a phenomena usually encountered in liquid chromatography-mass spectrometry (LC-MS) where ionisation efficiency of target analytes are altered in the presence of co-eluting compounds in the same matrix. It could cause either ion suppression or enhancement. • Quantitation of matrix effect is termed matrix factor (MF).	• Determination should be done by spiking at least 3 replicates of low (3 times LLOQ) and high (close to ULOQ) levels of concentration from at least 6 different sources/lots. • Use of fewer sources/lots may be acceptable in the case of rare matrices.	• Blank and zero calibrators should be free of interference at the retention times of the analyte(s) and IS. • IS response of the blank should not exceed 5% of average IS responses of calibrators and QCs. • Accuracy should be within ±15% and precision (%CV) should not exceed 15% in all matrix/slots.
Carryover	• Impact of carryover on the accuracy of study sample concentrations should be assessed.	• Should be assessed by injecting blank samples after a high concentration sample or highest calibration standard. • If unavoidable, study samples should not be randomised.	• Should not be greater than 20% of the LLOQ and 5% for the IS.
Parallelism	• Parallelism demonstrates that the serially diluted incurred sample response curve is parallel to the calibration curve. Parallelism is a performance characteristic that can detect potential matrix effects and interactions between critical reagents in an assay.	• A high concentration study sample (preferably close to Cmax) should be diluted to at least 3 concentrations with blank matrix. Two methodologies exist: • Method 1-linear regression analysis of the plotted measured concentrations of diluted samples against 1/dilution factor using log scales. • Method 2-plotting the back calculated concentration.	• Parallelism is improved when the linear fit shows a slope close to 1. • Precision between samples in a dilution series should not exceed 30%.

Validation parameter	Definition	Methodology	Acceptance criteria
Dilutional linearity/integrity	• Dilutional linearity demonstrates the accurate measurement of concentrations of spiked samples (i.e. QCs) exceeding the quantitation range when serially diluted to within the quantitative assay range.	• Spike blank matrix to concentrations above ULOQ and dilute down with blank matrix (5 determinations per dilution).	• Accuracy: ±15% of nominal concentrations. • Precision: ±15% CV.
Prozone (hook) effect	• The hook effect occurs when increasing analyte concentrations result in no change or decreased signals when compared to the preceding concentration.	• Spike the matrix with analyte concentration above ULOQ (undiluted) for hook effect and dilute sample (to at least 3 different dilution factors) with blank matrix to a concentration within the calibration range. • At least 3 runs for each dilution factor, the same number of replicates that were used for the study samples.	• Calculated concentration for each dilution should be within ±20% of the nominal concentration after correction for dilution, and the precision of the final concentrations across all the dilutions should not exceed 20%.

Abbreviations	
CC	Column Chromatography
CV	Coefficient of Variation
IS	Internal Standard
LBA	Ligand Binding Assay
LLOQ	Lower Limit of Quantification
LOD	Limit of Detection
LOQ	Limit of Quantification
MS	Mass Spectrometry
QC	Quality Control
RSD%	Relative Standard Deviation
ULOQ	Upper Limit of Quantification