Quantification in untargeted mass spectrometry-based metabolomics

Kloet, F.M. van der

2

ANALYTICAL ERROR REDUCTION USING SINGLE POINT CALIBRATION FOR ACCURATE AND PRECISE METABOLOMIC PHENOTYPING

ABSTRACT

Analytical errors caused by suboptimal performance of the chosen platform for a number of metabolites and instrumental drift are a major issue in large scale metabolomics studies. Especially for MS-based methods, which are gaining common ground within metabolomics, it is difficult to control the analytical data quality without the availability of suitable labeled internal standards and calibration standards even within one laboratory. In this paper we suggest a workflow for significant reduction of the analytical error using pooled calibration samples and multiple internal standard strategy. Between and within batch calibration techniques are applied and the analytical error is reduced significantly (increase of 25% of peaks with *RSD* lower than 20%) and does not hamper or interfere with statistical analysis of the final data.

Kloet, F.M. Van Der, Bobeldijk, I, Verheij, E.R and R.H. Jellema. Analytical Error Reduction Using Single Point Calibration for Accurate and Precise Metabolomic Phenotyping. Journal of Proteome Research. 2009 Nov;8(11):5132-41

2.1 INTRODUCTION

Recently there has been an explosion of analytical methods developed and applied in different metabolomics related research areas, such as nutrition research[90, 13, 135, 35] , drug discovery[63], optimization of fermentation processes [132, 131] and for breeding [24] of plants. In all these applications it is important to be able to understand and control factors that contribute to errors in the data and result in poor data quality. The total variation in a dataset is a function of different sources of variation[128]. The biological variation is present by design of the study and selection criteria of the subjects. In some cases, additional 'biological' variation can be introduced by differences in sample collection and sample storage [111, 25] Samples drawn from biological systems such as a microbiological fermentation or from body fluids like blood or urine are highly susceptible to changes due to biological reactions that take place, especially when the environment of the sample changes. It is therefore essential that changes in metabolites are minimized during sampling and sample preparation [131] in order to obtain a snap-shot representation of the biological system at the time of sampling. From an analytical point of view the data is of a much higher quality than years ago thanks to the efforts of instrument vendors to obtain more reproducible data, but it is still not enough. The analytical errors should be controlled as much as possible and reduced to a minimum and should not be confused with biological differences within the studied system.

2.1.1 *Sources of analytical variation*

A large part of the analytical variation is caused by suboptimal performance of the chosen platform for (sub-)sets of metabolites and instrumental drift. The ability of a method to detect a specific metabolite (i.e. its performance) is a complex interplay of its physical and chemical properties and is also partially dependent on the sample composition (matrix, e.g. ion suppression in MS based systems [11]) and in many cases on its concentration. Analytical variation for an individual analyte caused by differences in sample composition (matrix effect during extraction, derivatisation and analysis) can be removed only by using stable isotope labeled internal standards. The isotope labeled equivalent of the analyte performs the same as the original analyte and therefore differences in measured peak intensities for the isotope can directly be related to instrumental errors or sample preparation errors. Internal standards that are added before any sample preparation has been performed allow for correction of instrumental drift and sample preparation errors. Instrumental drift is especially important when a large number of samples are concerned, albeit within one batch or when they have to be divided over a number of batches. If instruments need to be cleaned within a series of measurements on samples of the same study, the data suffers from systematic differences between the batches. The severity of the systematic differences of course depends on the platforms being used and to a great extent on the chemical properties of the individual metabolites. Instrumental drift or offset between batches can be corrected for by internal standards or by calibration standards as is often done in bio-analysis or in targeted metabolite profiling [14]. In metabolomics

where hundreds of (also unidentified) metabolites are measured it is very uncommon to measure calibration standards for each of these metabolites. This would be very laborious and thus expensive, but equally important is the fact that beforehand it is not always clear which metabolite is of interest for the study at hand. In order to assess the data quality of all of these metabolites the use of pooled study (QC) samples has been described recently in the literature [28, 37]. In this approach pooled samples are analyzed regularly in between the individual study samples, several times within each batch. As a pooled study sample reflects the average metabolite concentrations within a study, this sample contains the same compounds (e.g. metabolites) as all the other samples. The performance of the analytical platform for all the compounds can be assessed by calculating the relative standard deviation (population standard deviation divided by the population mean) in these pooled samples [129, 28]. Various approaches to detect artifacts are described in Burton et al.[18]. Most of them however depend on visual inspection. Descriptions of data quality improvements are quite scarce. This paper describes a workflow for significant reduction of the analytical error using these pooled QC samples. Based on a multiple internal standard strategy and moreover between and within batch calibration techniques the analytical error is reduced significantly and will thus not hamper statistical analysis of the final data. Although this paper focuses on GC/LC-MS measurements and peak areas, it should be noted that the solution presented here is generic.

2.2 WORKFLOW AND METHODS

For an effective removal of different sources of analytical variation the preprocessing steps should follow a specific sequence. The first step is the data normalization using an internal standard. This step reduces the differences in sample extraction (which can be caused by slight differences in the composition of the samples) and also differences in the volumes injected. Especially the latter issue is of importance when injecting such low volumes as 1-2 μl. The second step is the removal of between-batch and within-batches batch offsets and drifts. This step can only be omitted if each metabolite has a structural analogue IS that corrects for all offsets and drifts, which is not the case in metabolomics analysis. The final steps consist of the combination of data from replicate sample analysis and removal of noise [13] and biomass correction [141]. The biomass correction neutralizes differences in response due to sample weight or volume (e.g. weighed liver tissue or dry matter within a suspension). As the biomass correction is a per-sample multiplicative correction it does not matter at which stage this is performed. The current paper focuses on the normalization using a single internal standard from a set of internal standards and the removal of between and even within batch differences by means of single point calibration using pooled quality control (QC) samples. These calibration techniques effectively render these QC samples, now used for calibration, useless for an independent assessment of the systems performance. To give an independent (unbiased) performance index it is suggested that in addition to these calibration samples additional QC samples should be measured as well which are used for validation of the results. The systematic approach of data pre-processing allowed the workflow to be implemented in a

fully automated environment.

2.2.1 *Used symbols and terminology*

Terminology

Internal standard: There are several definitions of internal standards and surrogates in the literature describing analytical methods. In this publication internal standard is a compound added to the sample before a critical step in the analysis. Depending on the method, the internal standard can be added before or during the extraction of the sample, derivatisation steps, etc.. An internal standard is not necessarily an isotope labeled version of an analyte but can also be structurally related to one or more analytes but not naturally occurring in the samples of interest. If a method covers analytes from different compound classes, multiple internal standards preferably covering all classes should be used.

Batch: a group of samples that has been extracted, derivatised (if applicable) and analysed together at the same time and using the same chemicals, same storage conditions.

QC sample: sample prepared by pooling aliquots of individual study samples, either all or a subset representative for the study. The QC sample has (should have) an identical or a very similar (bio) chemical diversity as the study samples. If insufficient sample volumes are available (e.g. rodent studies), samples collected outside the study but from a similar origin can be used. The QC samples are evenly distributed over all the batches and are extracted, derivatised (if applicable) and analysed at the same time as the individual study samples as part of the total sequence order.

QC calibration sample: sample chemically identical to the QC sample (from the same pool), prepared in the same way as QC samples. QC calibration samples are used for external calibration.

QC validation sample: sample chemically identical to the QC sample (from the same pool), prepared in the same way as QC samples. QC validation samples are solely used to monitor the result of all the data pre-processing steps and the quality of the full method. They are not used for external calibration.

Peak: For the purpose of this publication we use a broader definition of peak. A peak can be a single feature (intensity of a mass/ion or a different signal at a retention time or shift) or can be a sum of features (summed intensity of several ions at the same retention time). In our examples one peak represents one compound or metabolite detected in the data.

Analytical performance: The ability of an instrument to accurately detect a specific chemical component.

Assumptions

The method described in the procedure below focusses on peaks. The study samples vary in concentration for several peaks. For the purpose of this publication and all the procedures described here, we assume that response factors and the analytical performance of the internal standards and individual analytes are not influenced by the sample differences. In other words, the analytical performance of the method for all the individual analytes observed in the pooled QC sample will be the same in all other individual study samples.

Used symbols

i	index number for samples
p	Chromatographic peak (chemical component)
$C_{p,i}$	Concentration of peak p for sample i
F_p	Response factor for peak p
$F_p(t)$	Response factor for peak p at time point t
$F_{p,i}$	Response factor for peak p for sample i
$G_{p,i}$	Transformed form of $F_{p,i}$ after internal standard correction
$X_{p,i}$	Measured response of peak p for sample i
$X_{is,i}$	Measured response of internal standard peak is for sample i
$X'_{p,i}$	Relative response after internal standard calibration of peak p
$X'_{qc,p,b}$	Relative response after internal standard calibration of peak p for QC calibration samples in batch b
$X''_{p,i}$	Relative response after internal standard calibration and batch calibration of peak p for sample i
$A_{p,b}$	Average amplification relative response factor for peak p in batch b
$cf_{p,b}$	Calibration factor for peak p in batch b
$\alpha_{p,b,qc}$	Slope for linear estimate of QC calibration values for peak p in batch b
$\beta_{p,b,qc}$	Intercept for linear estimate of QC calibration values for peak p in batch b
$G_{p,b,i}$	Linear estimate of QC calibration values for peak p for sample i in batch b
Z	Smoothed estimate of QC calibration values for a single peak in a single batch

2.2.2 *Internal Standard normalization*

When focussing on MS analysis, it is generally difficult to model the extraction (derivatisation), MS ionisation and fragmentation variability of a compound by the behaviour of an internal standard with very different physical-chemical properties. This is especially the case when compound and reference belong to chemically different classes (e.g. glucose-d7 is a good representative for glucose but chances are high that it is not suitable for valine or a lipid). Theoretical and practical experiences indicate a positive effect from the use of a cocktail of stable isotope labelled internal standards with the same chemical diversity as the metabolites detected in the samples and exploit the close chem-

ical similarity (e.g. glucose-d7 is also expected to be a good internal standard for fructose and other hexoses). One could argue which internal standards, if more are included, should be used to adequately correct the errors in the measured responses of the individual metabolites. Sysi-Aho et al.[122] suggest selecting the best internal standard based on similarity between the distributions of the available internal standards and the compounds that are measured. Measurements that are performed at a later stage are then corrected using the preferred internal standard. This way however, real-time analytical variation is not included in the internal standard selection process which may result in suboptimal error correction. We suggest using analyte responses in quality control (QC) samples, regularly analyzed in between the study samples, as means to find the best internal standard. Using the relative standard deviations (*RSDs*) of the analyte response in the QC samples to quantify the amount of analytical variation, the best internal standard is the one that gives a minimum relative standard deviation.

In general, the response of a detector for a peak p can be defined as a product of its concentration C_p and a response factor F_p specific to this compound. For sample i the measured response $X_{p,i}$ is defined as shown in Equation 1.

$$X_{p,i} = C_{p,i} \cdot F_p \tag{1}$$

In an ideal situation F_p is constant and therefore measurements with a constant $C_{p,i}$ have identical responses. The *RSD* of QC samples for each peak would then be zero.
Internal standards are routinely used to correct systematic errors in the measured response by transforming the measured response $X_{p,i}$ into a relative response $X'_{p,i}$ using the measured response X_{is} of the internal standard is as denoted in Equation 2.

$$X'_{p,i} = \frac{X_{p,i}}{X_{is,i}} \tag{2}$$

This transformation and its error correcting effect is based on the assumption that for a perfect internal standard the sensitivity of the instrument for compound p is directly related to the sensitivity of the instrument for internal standard IS. In case of a non ideal standard, the corrective effect is not predictable. It may range from almost as good as the ideal internal standard to an actual increase of the error. In typical metabolomics methods the corrective effect of internal standards is highly variable because the number and the chemical diversity of the analytes exceed that of the internal standards (only a few metabolites form a perfect pair with a certain internal standard in a typical dataset).

The *RSD* for the QC samples is calculated using Equation 3, in which the standard deviation ($\sigma_{X'_{p,qc}}$) (after internal standard correction) is divided by the average (<>) relative response after internal standard correction($< X'_{p,qc} >$)

$$RSD_{p,qc} = \frac{\sigma_{X'_{p,qc}}}{< X'_{p,qc} >} \tag{3}$$

The best internal standard is the one that results in a minimal *RSD*. This *RSD* is calculated per peak. When measurements are divided over multiple batches the relative standard deviation is calculated over all QC samples.

2.3 BATCH CALIBRATION

Adjustments of the analytical instrument (e.g. maintenance, cleaning, tuning etc.) between batches of samples can be the cause of analytical errors that cannot be corrected for using solely internal standard calibration. This behaviour exerts itself in different response factors between and even within batches. We suggest that QC samples describe this type of analytical variation adequately and as a result, QC samples can be used as a means to correct for it. This type of correction is referred to as batch calibration.

2.3.1 *Mean and median correction (between-batch)*

Assuming that the measurement errors in a single batch are randomly distributed, then different batches can be compared and corrected using the average or median value of the QC samples in a batch. The average amplification relative response factor per peak per batch ($A_{p,b}$) can be written as the average (¡ ¿) of the responses after internal standard correction of the QC samples per batch (Equation 4).

$$A_{p,b} =< X'_{p,qc,b} > \tag{4}$$

The between batch calibration concerns the adjustment of the amplification factor $cf_{p,b}$ per peak with respect to a reference batch (in Equation 5 batch 1 is taken as reference).

$$cf_{p,b} = \frac{A_{p,1}}{A_{p,b}} \tag{5}$$

The error between measurements in a single batch is assumed to be a homoscedastic and random effect and therefore the same offset correction factors obtained from the QC samples can be transferred to the samples that are measured in between the different QC samples per batch.

$$X''_{p,b,i} = cf_{p,b} \cdot X'_{p,b,i} \tag{6}$$

As an alternative, the median can be used for determining the batch correction factor (in Equation 4) instead of the average. This has advantages over the mean in being a more robust measure. However, most parametric (statistical) tests, that for example facilitate outlier detection, are focussed on averages which makes the use of the average advantageous.

2.3.2 *Linear regression (within-batch)*

In many cases the response of the QC samples is not randomly distributed within a sequence of measurements and a notable drift can exist. In such cases the mean or median correction method will quite adequately correct for differences between batches but poorly for samples within a batch. If it is assumed that the behaviour between two consecutive QC samples is linear it can be modelled using first order regression. Mathematically, Equation 1 still holds but now F_p is not a constant factor but dependent on the time point

at which the sample was measured within a sequence. Equation 1 has to be rewritten to Equation 7.

$$X_{p,i}(t) = C_{p,i}(t) \cdot F_p(t) \tag{7}$$

Assuming that the analysis time of each sample is the same, time is equivalent to injection order and Equation 7 reduces to Equation 8.

$$X_{p,i} = C_{p,i} \cdot F_{p,i} \tag{8}$$

After internal standard normalization has been performed the definition of QC calibrated data follows Equation 9, in which $G_{p,i}$ is the transformed form of $F_{p,i}$ after internal standard correction. The data are not calibrated for between batch differences. This is done as the final step.

$$X''_{p,i} = cf_{p,i} \cdot X'_{p,i} = X'_{p,i} \cdot \frac{1}{G_{p,i}} \tag{9}$$

Because for each batch the correction factor is different, Equation 9 translates into Equation 10.

$$X''_{p,b,i} = cf_{p,b,i} \cdot X'_{p,b,i} = X'_{p,b,i} \cdot \frac{1}{G_{p,b,i}} \tag{10}$$

The estimated trend of the (relative) response for the QC samples per peak within a batch can be written as a function of injection order that is adjusted for slope β, and an intercept α (Equation 11).

$$X'_{p,b,qc} = \beta_{p,b,qc} \cdot i_{b,qc} + \alpha_{p,b,qc} \tag{11}$$

In case of a first order regression the factors α and β can be calculated using regular linear regression. Although higher order regression methods can be applied they heavily depend on the number of QC samples that are measured and are more sensitive to outliers. Using the regression coefficients from Equation 11 an estimate of the QC response can be calculated at each injection point i within a batch. In order to make a good estimation the QC samples should be distributed evenly within the measurements in a batch to ensure a good representation of the total drift during a batch and measured at each start and end of a batch to prevent extrapolation (errors).

$$G_{p,b,i} = \beta_{p,b,qc} \cdot i_b + \alpha_{p,b,qc} \tag{12}$$

Using this estimated trend, the relative response after internal standard calibration, per batch, is divided by this trend (Equation 13)

$$X''_{p,b,i} = cf_{p,b,i} \cdot X'_{p,b,i} = \frac{X'_{p,b,i}}{G_{p,b,i}} \equiv X''_{p,i} \tag{13}$$

2.3.3 *Linear smoother*

The assumption that the data between consecutive QC samples, within a batch, behave in the exact linear manner has a drawback if only a few QC samples measurement points are available or the QC samples exhibit too much

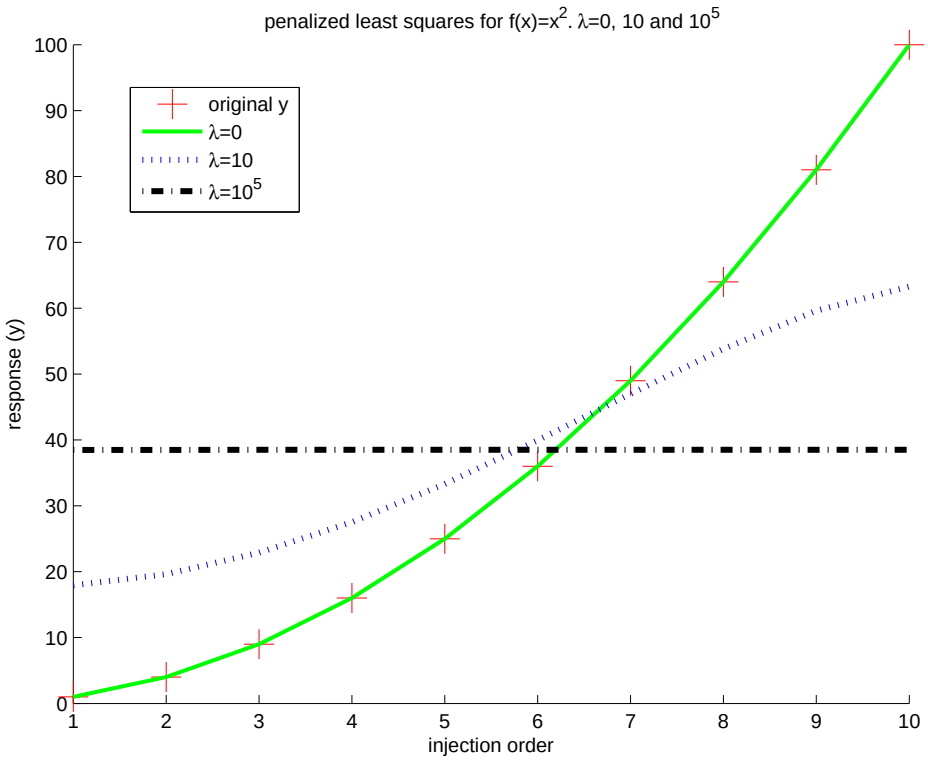


Figure 1: The effect of different values of λ on the smoothed estimate Z of an arbitrary trend exhibited by QC samples (e.g. $f(x) = x^2$). The black stippled line, where $\lambda = 10^5$ is used, is a horizontal line through the mean value of the QC samples. The blue dotted line, $\lambda = 10$, is a smoothed line that follows the general trend of the QC samples. The green line, $\lambda = 0$, follows the exact pattern of the QC samples.

variation (noisy data) for a significant linear trend. Linear correction would still improve the overall quality but could also introduce new analytical variation. In such cases the drift would best be described by a smoothed trend. Eilers[29] has shown that discrete penalized least squares can be used to estimate a smoothed trend. λ is the smoothing parameter where a larger λ results in a smoother estimate of the regression line Z. For really large values of λ, Z will result in a horizontal line (Figure 1, the black stippled line). This is a favourable characteristic because in cases of these large λs it is the only assumption that can be made (i.e. there's no overall linear relation between the QC samples). For small values of λ however, Z follows the trend exhibited by the QC samples (Figure 1, the blue dotted line). When no penalty is imposed the smoothed estimate Z follows the exact pattern of the QC samples (Figure 1 the green line).

To find an appropriate value for the penalty, λ_p has been made proportional to the residual error of the linear estimate ($G_{p,b}$) and the actual QC sample values ($X''_{p,b,qc}$) (Equations 11 and 12). Anything else than a perfect linear fit

results in a smoothed estimate of the trend between consecutive QC samples. The final QC trend is removed via

$$X''_{p,b,i} = \frac{X'_{p,b,i}}{Z'_{p,b,i}} \tag{14}$$

Finally the calibrated response is calibrated for between batch differences using the math as described in Equations 4 through 6. In this case however, $X'_{p,i}$ is substituted by $X''_{p,i}$.

2.4 EXPERIMENTAL

2.4.1 *Data sets*

To demonstrate the use of QC samples for the determination of the best internal standard and batch (between and within) calibration techniques two different datasets were used.
Data processing was performed using the MSD ChemStation E02.00.493 (Agilent technologies, Santa Clara, CA, USA). Based on many previous studies a target table is pre-defined containing matrix and study specific (metabolites with known and unknown identities). Each peak is characterized by its retention time and selected specific m/z value. For each study an update of the retention times (and in limited cases m/z) values is prepared. A limited number of selected chromatograms are compared to a chromatogram of a reference sample (sample that is analysed in each study). Peaks that have not been observed previously are added to the target table. Artefacts of the method are removed from the data as well as (multiple) entries for a single (identified) metabolite caused by different derivatization products of which the performance is known to be irregular. For plasma, typically 120-200 metabolites are reported. Even though this procedure is quite time-consuming, we believe it gives more reliable data than peak picking procedures or most deconvolution procedures with less missing values and less peaks. During acquisition of both datasets the following compounds were used as internal standards: Alanine d4 (ALA-D4), Cholic acid d4 (CA-D4), Leucine d3 (LEU-D3), Phenylalanine d5 (PHE-D5), Glutamic acid d3 (GLU-D3), Dicyclohexylphtalate (DCHP), Difluorobiphenyl (DFB), Trifluoroacetylantracene (TFAA). All compounds were purchased from Sigma (Zwijndrecht, the Netherlands).

Example metabolomics study 1

A nutritional intervention study that involved 36 volunteers. Volunteers were divided into 4 groups and received 4 different treatments: A, B, C and D, including placebo. A cross-over design was used in this study, with each group receiving each of the treatments, in a randomized order [7]. At the end of each treatment period each subject received an oral lipid challenge test, after which several blood samples were collected. Plasma samples collected within this study were analysed using different metabolomic platforms including the GC-MS method as described by Koek et al.[72] From the challenge test, only samples from treatment groups A and B were analysed by GC-MS

and this data is used for the application demonstration of the developed workflow and methods. Plasma samples (100 μl) were extracted with methanol and after evaporation the metabolites were derivatized (oximation and silylation). 8 different internal standards were added to the samples before the different sample preparation steps. The number of individual study samples analysed by GC-MS was 504. The samples were analysed in 18 batches, each batch contained 28 study samples (all timepoints from two subjects, randomized per subject) and 3 pooled QC samples. Each study sample was injected once; each QC sample was injected twice per batch. The QC injections were distributed evenly in the batch: at the start of the batch, after approximately every 6 samples and at the end of each batch. Besides these QC samples, additional QC validation samples were included. Each batch contained 1 QC validation sample. At the end of the analysis, a 19th batch was included, which contained a real replicate analysis of the complete time profile of two selected subjects. For this purpose a separate aliquot of the samples was extracted, derivatised and analysed. Data processing was performed as described above. 145 peaks (excluding internal standards) were reported.

Example metabolomics study 2

An inflammation modulation study with placebo and diclofenac was performed in parallel [144]. Each group had 10 volunteers. 19 volunteers completed the treatment. Blood samples were taken after an overnight fast on days 0, 2, 4, 7 and 9. Subjects underwent an oral glucose tolerance test (OGTT) on day 0 and day 9 of the study. Blood samples were taken just before (0 minutes) and 15, 30, 45, 60, 90, 120 and 180 minutes after the administration of the glucose solution (75 grams). The samples taken at day 9 were analysed using the same analytical methods described for example study 1. The number of individual study samples analysed by GC-MS was 361 (19 volunteers, 19 timepoints per volunteer), each sample was analysed twice, resulting in 722 sample injections. The samples were analysed in 26 batches, each batch contained 35 injections including QC samples. Batch 1 started with all samples of one of the subjects, timepoints randomized, followed by the randomized replicate measurements of the same subject until the maximum of 29 (sample) injections was reached. The next batch started with the remaining (replicate) samples of the previous subject followed by the, timepoint randomized, full set of samples of the next subject etc. In this way for each subject at least one replicate of the full time profile was analysed within one batch. The QC injections were distributed evenly in the batch: at the start of the batch, after approximately every 6 sample injections and at the end of each batch. No additional QC validation samples were measured. To assess the effect of the different preprocessing steps the replicated measurements were used. Data processing was performed as described above. 137 peaks (excluding internal standards) were reported.

Data extraction

Both example studies were processed using a target approach. The target table was adjusted 3 times for retention time shifts caused by shortening the column by several centimeters after each batch.

Data processing

Prototyping of the correction methods was executed in Matlab version 2007b[50]. The final implementation of the software was done in SAS version 9.1.3[52] as stored procedures complementary to a data warehouse (SAS) in which the study data were captured.

2.4.2 *Results and discussion*

Internal Standard normalization

The number of internal standards is dependent on the analytical method with a minimum of 1 and no maximum. To mimic the behaviour of the analytes structure analogues and stable isotope labelled compounds were used added as internal standards. In our example study we used 8 internal standards and applied our selection method to select the most suitable one. Table 1 shows the results of the selection method on the *RSD* values of the QC calibration samples. That the *RSD* indeed seems to be functioning as a good criterion for the selection of the best internal standard is shown in Table 2 in which examples are shown of the internal standard that was selected as best for a number of identified metabolites in metabolomics study 1. In all cases where an analyte had an own deuterated internal standard, this standard was selected, as it also gives the lowest *RSD* in the QC validation samples. For compounds with no deuterated analogue a structurally related IS was selected, e.g. LEU-D3 was selected for the corrections of both leucine and isoleucine, for alanine, ALA-D4 is selected. For some aminoacids within this study, the the deuterated analogue gives a slightly higher *RSD* than a different deuterated IS (8% vs 5%). Within this study this is the case for glutamate, for this compound PHE-D5 is selected as the IS giving the lowest *RSD* in the QC validation samples. Within this study, not all the reported metabolites have a suitable IS structurally. For example the detected fatty acids are less volatile and elute later in the chromatogram. For these matabolites the described procedure chooses apolar and/or late eluting internal standards such as DFB or DCHP as the most suitable. It should be noted that each metabolite might have several suitable IS within the approach giving similar *RSD* values. Therefore the chosen IS can differ from study to study, depending on the dataset.

RSD	Corrected for 1 IS (DCHP)	corrected for best IS
0%-10%	26%	58%
10%-20%	32%	24%
20%-30%	27%	10%
>30%	16%	8%

Table 1: Frequency distribution of *RSD* values of QC calibration samples from the first example study. The effect of the 'best' internal standard clearly translates into more peaks with lower *RSD* values.

Each of the peaks was assigned a best internal standard (1 out of 8) using the criterion as described in Equation 3. After the normalization step with the appropriate internal standard, the PCA score plot (Figure 2) shows cluster-

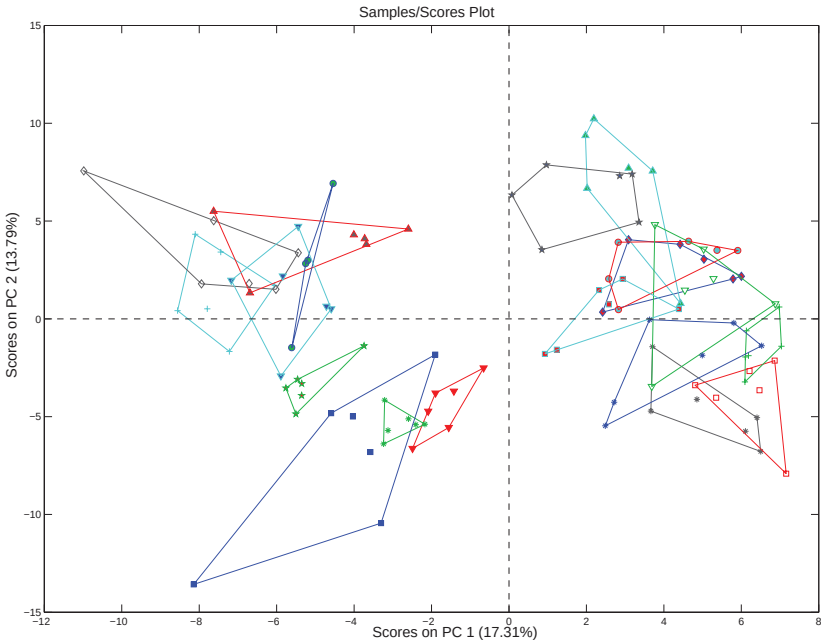


Figure 2: PCA score plot of QC calibration samples from the first example study after internal standard normalization. Different colours refer to the different batches. The data are autoscaled. Two clusters are visible; one before cleaning the MS source (left) and one after the cleaning procedure (right).

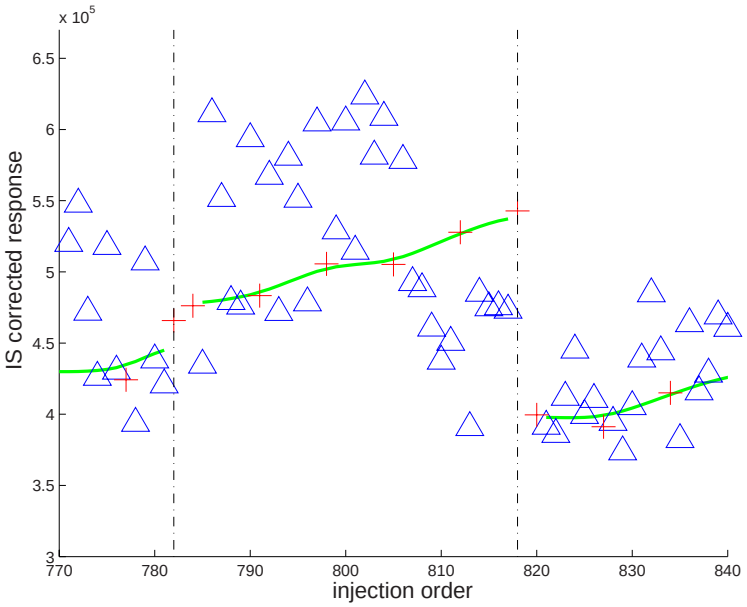
ing of the different QC samples per batch. Furthermore, the plot shows that there's a significant difference between two clusters of batches. Investigation reveals that the group on the left hand side are batches 1-9 whilst the remaining batches 10-19 are plotted on the right hand side; it coincides with the fact that the MS source was cleaned after the 9th batch. This behaviour emphasizes that QC samples (indeed) characterize the systems state (or can be used to do so). It also leads to the conclusion that remaining variation due to between-batch and/or within-batch differences is insufficiently corrected for using the normalization procedure with internal standards.

Metabolite	IS selected as best
Alanine	ALA-D4
Leucine	LEU-D3
Isoleucine	LEU-D3
Glutamic acid	PHE-D5
C16:1 Fatty acid	DFB
C16:0 Fatty acid	DCHP
C17:0 Fatty acid	DCHP

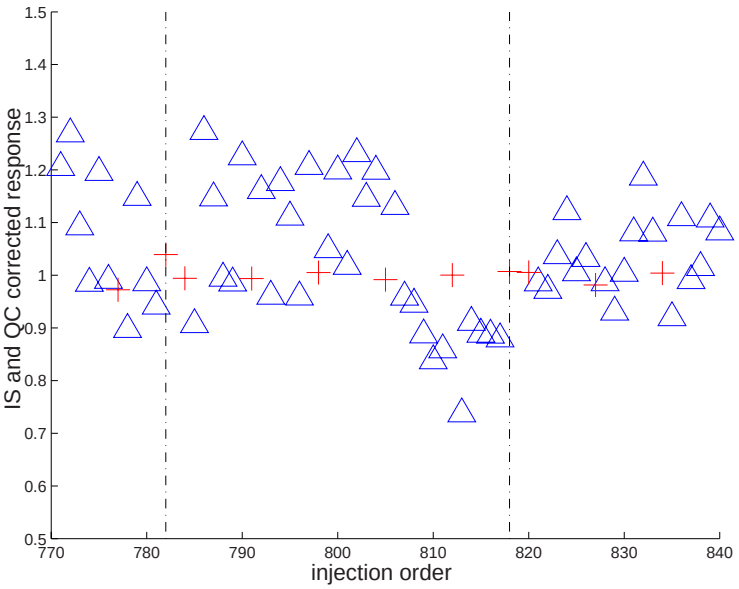
Table 2: Examples of the selected best internal standard for a selection of metabolites found in the first example study.

Batch calibration

Figure 3a shows a part of a time order profile plot of a specific metabolite from the second example study. The real samples are represented by the blue triangles and the QC samples by the red crosses, the green lines represent the smoothed estimate for the trend exhibited by the QC samples. Each vertical dashed line represents the start of a new batch. The figure clearly shows the effect of analytical errors that are introduced during measurements. Even though the data was corrected for the best internal standard, for some metabolites such as this example, large between-batch and within-batch differences still exist for the QC samples. The study was not set up for quantification purposes but it is apparent that for any further (statistical) analysis this type of error should be removed. In order to do so, a smoothed trend, per batch, was fitted through the QC samples. Figure 3b shows the results of the same metabolite after the batch calibration step. The *RSD* value of this metabolite for the QC samples dropped from 13.7% after internal standard correction to 2.1% after the additional batch correction step. The QC samples follow an almost horizontal line indicating that the within batch calibration was applied successfully. Furthermore, it also shows that the offset differences between the batches have been removed. The resulting variation is mainly due to the actual compositional differences between the samples (different subjects and different timepoints).



a



b

Figure 3: A part of the time order plot of a specific metabolite from the second example study after internal standard normalization (a) and batch calibration (b). The study samples are represented by triangles and the QC samples by crosses. The vertical dashed lines represent the start/end of a batch. The green lines represent the smoothed estimate of the QC samples. Fig. a shows the data after best internal standard normalization. Offset differences and within batch trends are clearly visible. Fig. b shows the same data but now after additional batch calibration. The QC samples clearly follow a horizontal line and no offset differences between batches are visible.

Validating the results

The availability of QC calibration, QC validation samples and replicated sample measurements allow for a triple validation check. The performance statistics used are as follows:

1. Improvement of *RSD* in the QC calibration samples
2. Improvement of *RSD* in the (independent) QC validation samples
3. Improvement of differences between samples with different composition (representative replicates)

The effect induced by the different calibration steps on the *RSD* value of the QC validation samples from the first example study is shown in Table 3. The table shows the distribution of the number of metabolites when the *RSD* range is divided into 4 classes. For replicated measurements the results are shown in Table 4. The results in Table 3 and Table 4 are comparable, the removal of between and within batch differences using the real-time variation information embedded within pooled QC samples shows a significant impromevent in observed *RSD* values.

RSD	Raw data	IS calibrated	IS + batch calibrated
0%-10%	12%	58%	81%
10%-20%	59%	24%	16%
20%-30%	16%	10%	3%
>30%	13%	8%	0%

Table 3: Frequency distribution of *RSD* values of the QC validation samples from the first example study. The IS normalization and batch calibration steps clearly have a favourable effect on the *RSD* frequency distribution of these samples.

RSD	Raw data	IS calibrated	IS + batch calibrated
0%-10%	49%	53%	72%
10%-20%	36%	32%	21%
20%-30%	8%	8%	7%
>30%	8%	7%	1%

Table 4: Frequency distribution of *RSD* values of duplicated measurements of real samples from the second example study. The calibration steps show the same positive effect on the frequency distribution of '*RSD*'s of these replicate measurements as the QC validation samples.

Visualizing information potential

To get a global idea about the information potential within the study data, a scatter plot of the analytical performance versus its reproducibility for all metabolites in a given data set is a remarkably elegant and simple method for

depicting the improved data quality obtained with the QC calibration method. An example for GC-MS data from the second example study is shown in an Information Density plot (ID plot) (Figure 4 a and b). The x-axis shows the *RSD*, and the y-axis shows the correlation between replicates. The correlation is both a measure of data quality and the range of observations. For example, a metabolite with a relatively large *RSD* (e.g. 50%) and a large concentration range (e.g. one order of magnitude between lowest and highest) typically has a good replicate correlation. On the other hand a metabolite with an excellent *RSD* (e.g. <10%) may have a very poor replicate correlation if its concentration in all samples is identical.

These plots visualize the proportion of good quality data (left of the vertical line / limit) and the proportion of metabolites with a wide concentration range (above the horizontal line / limit). The upper-left corner contains the most information rich metabolite data and it is obvious that the QC correction indeed shifts metabolites towards the lower (good) *RSD* region and results in more metabolites in the upper left hand corner region. Similar results are obtained if the concentration range is used instead of the replicate correlation coefficient. Depending on the objective of a metabolomics study these plots may be used in various ways for variable selection prior to statistical analysis of the data.

It is important to understand that the variability of the QC samples should represent the variability of the study samples. Therefore, the QC samples should be handled as if they were study samples which means for example that reuse of 'old', already extracted and derivatised QC samples is not a very good idea because it will induce an extra source of variation that is not present in the study samples. For large, long duration studies we suggest preparing sufficient number of QC aliquots, such that an identical sample is used throughout the whole study. Hereby we assume that the influence of the storage of the QC sample over longer periods of time has a negligible effect on the composition of the sample

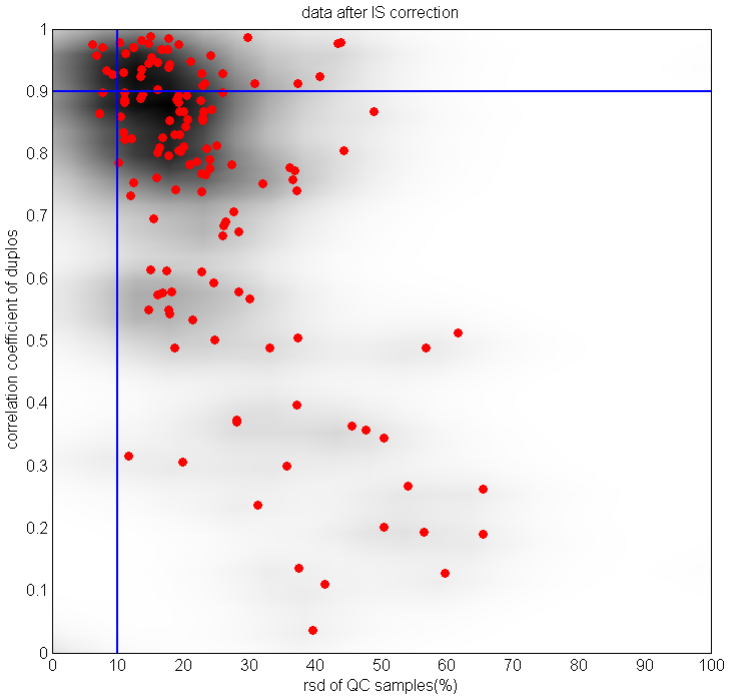
2.4.3 *Recommendations*

Internal Standard normalization

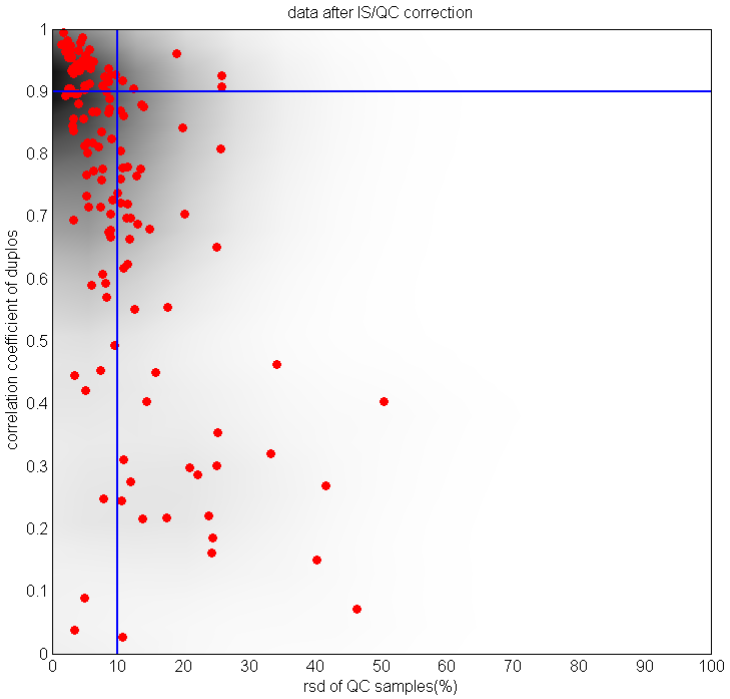
The most important descriptors of data quality are accuracy and precision. Standards and calibration curves are typically not used in unbiased metabolomics. This makes it impossible to assess the accuracy of the method for all the metabolites measured, identified and unidentified, and as a consequence it is equally impossible to improve the accuracy for one or more metabolites. The described procedure focuses on improving data precision, which is equivalent to minimizing the *RSD*. From the point of view of optimising data quality it would be beneficial to use a cocktail/mix of deuterated internal standards that have structures that are analogue to the ones that have to be analyzed.

Batch calibration

QC samples that are generally used to assess the performance of the system are now used for calibration purposes. To obtain a calibration model, at least 2 QC calibration samples should be measured per batch, one at the beginning and one at the end of each batch. To increase the robustness and reliability of



a



b

Figure 4: Information Density plot (ID plot), a scatter plot of the analytical performance versus its reproducibility for all metabolites for data from the second example study. Fig a. shows the results after internal standard normalization, Fig b. shows the results after internal standard normalization and batch calibration.

such a model more QC calibration samples should be measured per batch. The actual number of QC samples depends on the individual analytical method, its robustness and performance characteristics and the availability of QC material. There should be a balance between the number of QC samples and the number of study samples in order to keep the analysis efficient and low cost. The number of replicate injections per QC sample should be equal or similar to the number of injections per individual study sample. It is also clear that outliers in QC samples can adversely affect the outcome.

2.5 CONCLUSION

Results from metabolomics studies can be improved using a single point calibration based upon results obtained from pooled QC samples that are repeatedly measured in between study samples. This one point calibration is indispensable when large scale metabolomics studies are performed and both within and between batch differences become a problem due to instrumental and environmental changes during measurements. We have shown that two types of QC samples are required whereby the first type, calibration QC samples, is used to perform a one point calibration, and the second type, validation QC samples, is used to assess how well the calibration procedure improved the data quality. We have shown that it is feasible to increase the number of metabolites with a relative standard deviation for replicated measurements below 10% from 49% of the peaks to 72%. As a result, the induced or biological variation in the study samples becomes more apparent and more meaningful statistical models can be build from the corrected data.
We have also shown that the *RSD* in the QC samples before and after internal standard correction is a good measure to find the best match between a given metabolite and the set of internal standards that were spiked in the sample. This is a practical alternative to using a separate (isotope labeled) internal standard for each metabolite, which is not feasible due to costs and availability. For large data sets, it is a difficult task to obtain an idea about the information content of the data and to make a comparison between a before and after correction situation. For this purpose, we suggest and demonstrate a new method for presentation of the total dataset focusing on the analytical variability (*RSD*) and the concentration or intensity range of the metabolites. The example shown in this paper clearly shows an improved information content after correction of the raw data with internal standards and QC samples.

The methodology presented here was applied to GC-MS data but is applicable also to other datasets obtained with other analytical techniques. A future perspective for further development of the methodology lies in the fusion of datasets obtained from different studies or different instruments. A copy of the MATLAB[50] prototyping code is available on request from the corresponding author.

ACKNOWLEDGEMENT

The authors would like to acknowledge Bas Muilwijk and Marc Tienstra for the GCMS data. The help of Gertruud Bakker and Carina Rubingh with the data analysis and randomization schemes is acknowledged.