Anal. Chem. **2010,** *82,* 1039–1046

Equating, or Correction for Between-Block Effects with Application to Body Fluid LC−MS and NMR Metabolomics Data Sets

Harmen H. M. Draisma,[†] Theo H. Reijmers,[†] Frans van der Kloet,[†] Ivana Bobeldijk-Pastorova,[‡] Elly Spies-Faber,[‡] Jack T. W. E. Vogels,[‡] Jacqueline J. Meulman,[§] Dorret I. Boomsma,[||] Jan van der Greef,[†] and Thomas Hankemeier*[,†]

Leiden/Amsterdam Center for Drug Research (LACDR), Leiden University, P.O. Box 9502, NL-2300 RA Leiden, The Netherlands, TNO Quality of Life, P.O. Box 360, NL-3700 AJ Zeist, The Netherlands, Mathematical Institute, Leiden University, P.O. Box 9512, NL-2300 RA Leiden, The Netherlands, and Department of Biological Psychology, VU University Amsterdam, Van der Boechorststraat 1, NL-1081 BT Amsterdam, The Netherlands

Combination of data sets from different objects (for example, from two groups of healthy volunteers from the same population) that were measured on a common set of variables (for example, metabolites or peptides) is desirable for statistical analysis in "omics" studies because it increases *power*. However, this type of combination is not directly possible if nonbiological systematic differences exist among the individual data sets, or "blocks". Such differences can, for example, be due to small analytical changes that are likely to accumulate over large time intervals between blocks of measurements. In this article we present a data transformation method, that we will refer to as "quantile equating", which per variable corrects for linear and nonlinear differences in distribution among blocks of semiquantitative data obtained with the same analytical method. We demonstrate the successful application of the quantile equating method to data obtained on two typical metabolomics platforms, i.e., liquid chromatography−mass spectrometry and nuclear magnetic resonance spectroscopy. We suggest uni- and multivariate methods to evaluate similarities and differences among data blocks before and after quantile equating. In conclusion, we have developed a method to correct for nonbiological systematic differences among semiquantitative data blocks and have demonstrated its successful application to metabolomics data sets.

Combining data from different sources is an important topic in systems biology. At least two types of data combination can be envisaged. The first type of combination is often referred to as data integration or data fusion, and here combination is considered of data sets all representing the same set of objects (for example, a group of healthy volunteers) but different sets of measured variables (for example, metabolites, peptides, etc.).[1,2] Data fusion combines the strengths of different analytical techniques to enhance the biological interpretation of the variability present in the study population. In the second type of combination, which is the scope of this article, data sets are combined representing different groups of objects (for example, two groups of healthy volunteers) that were measured on a common set of attributes (for example, the same set of metabolites). Combination of data sets in such a way is desired because it increases the *power* of statistical analyses. In other words, one may want to combine different data "blocks". In this article, we use the term "blocks" to refer to measurements obtained on the same analytical method but on different sets of objects and in particular with a considerable time span in between these sets of measurements. A block can consist of data from one or more measurement batches. A similar definition of blocks is given by Zelena et al.[3]

Different measurement blocks can arise within a study, for example, because (1) the number of study samples is too large to measure all samples in one measurement block or in one laboratory, (2) additional samples become available in the course of the study while previously collected samples have already been measured, or (3) following a successful pilot experiment, additional samples are measured for validation. It is also conceivable that it is desired to combine data blocks from different studies. Nonbiological differences between the data from different measurement blocks can exist due to small analytical differences that are often unavoidable and that are typically not addressed during method robustness tests. Such analytical differences are, for example, likely to accumulate over large time spans between blocks of measurements.[3−5]

* To whom correspondence should be addressed. E-mail: hankemeier@chem.leidenuniv.nl. Fax: +31-71-527-4277.

[†] LACDR, Leiden University.

[‡] TNO Quality of Life.

[§] Mathematical Institute, Leiden University.

[||] VU University Amsterdam.

(1) Steinmetz, V.; Sévilla, F.; Bellon-Maurel, V. *J. Agric. Eng. Res.* **1999**, *74*, 21–31.

(2) Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van der Vat, B. J.; Jellema, R. H. *Anal. Chem.* **2005**, *77*, 6729–6736.

(3) Zelena, E.; Dunn, W. B.; Broadhurst, D.; Francis-McIntyre, S.; Carroll, K. M.; Begley, P.; O'Hagan, S.; Knowles, J. D.; Halsall, A.; Wilson, I. D.; Kell, D. B. *Anal. Chem.* **2009**, *81*, 1357–1364.

(4) Feudale, R. N.; Woody, N. A.; Tan, H.; Myles, A. J.; Brown, S. D.; Ferré, J. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 181–192.

(5) Alam, T. M.; Alam, M. K.; McIntyre, S. K.; Volk, D. E.; Neerathilingam, M.; Luxon, B. A. *Anal. Chem.* **2009**, *81*, 4433–4443.

10.1021/ac902346a
Published on Web 01/07/2010

In data fusion, often three types of combination of data from a common set of objects are considered: high-level fusion, which is the combination of results of data analyses obtained on sets of different variables, low-level fusion, or the concatenation and possibly subsequent weighting of data matrices in such a way that the objects are the shared mode, and mid-level fusion, a term used to describe the combination of variables selected from different data sets.[1,2] A similar classification can be envisioned when considering combination of data on sets of different objects where the attributes are identical. In this article, we present a method that enables such combination of data blocks at a "low level" and illustrate its use with metabolomics data sets. Combination at low level allows maximal flexibility in the choice of subsequently applied (multivariate) data analysis methods yielding results for the combined data sets and therefore is particularly suited to increase the *power* of such subsequent data analyses. Moreover, combination of data at a low level allows to account for differences in distribution shapes of the same variable(s) among the data sets to be combined, if it is known that such differences have a nonbiological cause.

The necessity and possibility of applying data correction methods in order to obtain combinable "omics" data blocks will vary from situation to situation. In the discussion below, we have intended to provide a guideline where we start with a description of situations where combination should be possible without additional data correction and end with a description of situations where the data transformation method we propose in this article could be useful.

1. If the between-block reproducibility of the used analytical method is good (e.g., semiquantitative nuclear magnetic resonance (NMR) spectroscopy under similar conditions for all measurement blocks of which data sets are to be combined),[6,7] or the data sets to be combined all contain quantitative data (either through separate calibration per measurement block or through transfer of calibration models),[4,5] then the combination of data sets from different measurement blocks should be possible without additional correction. However, currently obtaining quantitative data from metabolomics experiments is still rather difficult, because often due to the absence of reference standards for all detected compounds it is impossible to create a complete calibration model per variable.[8] Both techniques that are the most frequently used in metabolomics, i.e., liquid chromatography−mass spectrometry (LC−MS) and NMR, suffer from this problem.
2. If the measurements performed within particular blocks are not reliable, then the data from these measurements should be discarded. The reliability of measurements can be monitored using, for example, a quality control (QC) sample consisting of pooled individual study samples, of which aliquots are measured during all analytical measurement blocks.[8−13]
3. Recently, a method has been presented to correct for between-batch effects using these repeated measurements of QC samples as well.[14] Like the other methods to be discussed below, it can be used for the correction of semiquantitative data, i.e., in cases where no full calibration models can be made. We will refer to techniques that make combinable sets of semiquantitative data as "equating" methods, because the term "equating" is used in psychometrics to denote techniques that solve similar problems.[15,16] In the method of van der Kloet et al., the data are corrected for within-batch and between-batch effects per metabolite using the responses of pooled QC samples (for that metabolite).[14] This method can be of use if a single-point calibration is appropriate for correcting differences in data distributions among measurement batches or even among measurement blocks. Of course, it can be used only if the same QC samples are measured in all batches or blocks of which data need to be combined.
4. There are situations where repeated QC sample measurements cannot be used for between-batch effect correction or for between-block effect correction. An obvious example is if such measurements have not been done during all measurement batches or blocks of which data sets need to be combined. Another example is when the QC samples are not representative for the measurements in all data sets to be combined. This can happen for instance if there is differential degradation in the QC samples with respect to the individual study samples. Such situations are analogous to the situations where in the context of multivariate calibration transfer one would typically use "nonstandardization methods", i.e., data preprocessing methods that are independent of transfer standards.[4] An example of an equating method that is independent of repeated QC sample measurements is local autoscaling: autoscaling per data set separately.[17] Like the method described in ref 14, this local autoscaling method could be regarded as a linear equating method.
5. Finally, the data distribution shapes of the same variable in all data sets to be combined can be different mainly due to nonbiological differences among the blocks. Such nonlinear differences among the data distribution shapes in different blocks can arise even if within each block

(6) Keun, H. C.; Ebbels, T. M.; Antti, H.; Bollard, M. E.; Beckonert, O.; Schlotterbeck, G.; Senn, H.; Niederhauser, U.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Chem. Res. Toxicol.* **2002**, *15*, 1380–1386.

(7) Dumas, M. E.; Maibaum, E. C.; Teague, C.; Ueshima, H.; Zhou, B.; Lindon, J. C.; Nicholson, J. K.; Stamler, J.; Elliott, P.; Chan, Q.; Holmes, E. *Anal. Chem.* **2006**, *78*, 2199–2208.

(8) Sangster, T.; Major, H.; Plumb, R.; Wilson, A. J.; Wilson, I. D. *Analyst* **2006**, *131*, 1075–1078.

(9) Gika, H. G.; Theodoridis, G. A.; Wingate, J. E.; Wilson, I. D. *J. Proteome Res.* **2007**, *6*, 3291–3303.

(10) Theodoridis, G. A.; Gika, H. G.; Wilson, I. D. *Trends Anal. Chem.* **2008**, *27*, 251–260.

(11) Burton, L.; Ivosev, G.; Tate, S.; Impey, G.; Wingate, J.; Bonner, R. *J. Chromatogr., B* **2008**, *871*, 227–235.

(12) Dunn, W. B.; Broadhurst, D.; Brown, M.; Baker, P. N.; Redman, C. W. G.; Kenny, L. C.; Kell, D. B. *J. Chromatogr., B* **2008**, *871*, 288–298.

(13) Bijlsma, S.; Bobeldijk, I.; Verheij, E. R.; Ramaker, R.; Kochhar, S.; Macdonald, I. A.; van Ommen, B.; Smilde, A. K. *Anal. Chem.* **2006**, *78*, 567–574.

(14) van der Kloet, F. M.; Jellema, R. H.; Verheij, E. R.; Bobeldijk, I. *J. Proteome Res.* **2009**, 5132–5141.

(15) Kolen, M. J.; Jarjoura, D. *Psychometrika* **1987**, *52*, 43–59.

(16) Van der Linden, W. J. *Psychometrika* **2000**, *65*, 437–456.

(17) Wagner, S.; Scholz, K.; Sieber, M.; Kellert, M.; Voelkel, W. *Anal. Chem.* **2007**, *79*, 2918–2926.

the measurements for each variable are within the dynamic range of the detector. For example, in case of LC−MS, in a typical metabolomics study, measurement values can be outside the linear range for various reasons: saturation of the detector, peak integration effects (e.g., caused by peak tailing, depending on the concentrations of a particular compound in the samples measured in a particular block), or nonlinear losses during sample preparation. These effects can be different for different measurement blocks. In this article, we propose an equating method that corrects for nonlinear differences between distributions under the assumption that there is an underlying common distribution. Therefore, the beneficial effects of our method will be largest when the compositions of the object groups are balanced among the measurement blocks of which data are to be combined. Our method is independent of repeatedly measured QC samples as well.

In case it has been decided that equating methods need to be considered to correct the data for between-block effects, the choice of a particular equating method might not be trivial. It can be generally stated that the equating method should be used that removes most analytical between-block variation with respect to the biological variation present in all blocks. In practice, however, it is not always possible to determine exactly which part of the total between-block variation is attributable to biological variation and which part is attributable to analytical variation, because the objects measured in different blocks are different. In this respect, an objective evaluation of the results of equating is necessary, because the best equating method in a given situation is not necessarily the one that gives the most desirable results in view of the biological question. Therefore, as with any data preprocessing, using the results of subsequent data analyses alone as a reference to "optimize" the choice for a particular method could lead to bias.

The structure of the remainder of this article is as follows. In the Materials and Methods section, we first introduce the metabolomics data that we will use to illustrate the use of our equating method. Then, we describe our equating method. Univariate as well as multivariate parameters are described that can be used to evaluate the comparability of data sets before and after equating. The Results and Discussion section describes the results of application of our equating method to the data sets originating from the different measurement blocks. Several possible sources of nonbiological systematic variation between data obtained in the different blocks are pointed out.

The results of application of our equating procedure to metabolomics data sets, as described in this article, will be used to reproduce and extend our observations that were done in a cohort of twins.[18] The results of these subsequent analyses on the combined equated data sets described in the current article will be presented in a separate paper, because the biological interpretation of the results is out of the scope of this paper.

MATERIALS AND METHODS

Participant recruitment and characterization, blood sampling, and blood plasma sample preparation were performed as described previously.[18] In brief, blood was drawn and urine collected from all participants (twins and biological nontwin siblings) after overnight fasting. Plasma samples were stored at −80 °C until analysis. The LC−MS and ^{1}H NMR measurements were performed in two blocks; the measurements of "block 2" (B2) were performed almost 1 year (48 weeks) after those of "block 1" (B1). In B2, for the purpose of QC of the LC−MS and NMR analyses, QC samples were prepared prior to sample preparation by pooling equal amounts of plasma sample from all participants who were measured in that block. In B1, such QC samples were prepared for the LC−MS analyses only. For both LC−MS and NMR analyses, these QC samples were inserted uniformly distributed after separate randomization of the measurement order of the individual study samples in each batch.

LC−MS Plasma Lipid Profiling. Plasma lipid extraction and profiling by LC−MS were performed as described previously.[18] After lipid extraction, all extracts were stored at −20 °C and measured within 2 weeks. Each peak area obtained for a lipid was corrected using an appropriate internal standard (IS), which had been added prior to sample preparation; no further normalization of the data was applied.

^{1}H NMR Analysis of Plasma. Prior to ^{1}H NMR spectroscopic analysis, 300 μL of each plasma sample was centrifuged to remove proteins that had come out of the solution after freezing and transferred to a 5 mm o.d. NMR tube. To each sample 300 μL of deuterated sodium phosphate buffer (0.1 mmol/L, pH 7.4, made up with D_2O) was added. ^{1}H NMR spectra were acquired in triplicate on a fully automated Bruker Avance 600 MHz spectrometer (Bruker Analytik GmbH, Karlsruhe, Germany) using a "Carr−Purcell−Meiboom−Gill" (CPMG) spin−echo pulse sequence and operating at an internal probe temperature of 300 K. The water signal was removed by a presaturation technique in which the water peak was irradiated with a constant frequency during the relaxation delay. A total of 128 transients were acquired into 32×10^3 data points for B1 and 64×10^3 data points for B2. A spectral width of 6 kHz for B1 and 12 kHz for B2 was used with a spin relaxation delay of 88 ms and τ 3.4×10^{-4} s for both blocks. The spectra were processed using XWIN-NMR software (v.3.1, Bruker Analytik GmbH). An exponential line-broadening function of 0.5 Hz was applied to the free induction decays (FIDs) prior to Fourier transformation. All spectra were manually phased, baseline-corrected, and referenced to the lactate signal (CH_3 δ 1.33). After peak picking of the NMR data using the XWIN-NMR software, peak lists were imported into Winlin (V1.10, TNO, The Netherlands). Small variations in chemical shifts in the NMR spectra were adjusted manually based on the partial linear fit algorithm.[19] The peak-picked data from B1 and B2 were aligned together, with the aim to make the alignment for data from both blocks as comparable as possible. Peaks detected in

(18) Draisma, H. H.; Reijmers, T. H.; Bobeldijk-Pastorova, I.; Meulman, J. J.; Estourgie-Van Burk, G. F.; Bartels, M.; Ramaker, R.; Van der Greef, J.; Boomsma, D. I.; Hankemeier, T. *Omics* **2008**, *12*, 17–31.

(19) Vogels, J. T. W. E.; Tas, A. C.; Venekamp, J.; Van der Greef, J. *J. Chemom.* **1996**, *10*, 425–438.

at least 80% of the spectra recorded in each block were kept for further analysis.[2,13] Then, the data were median-normalized.[20]

Differences between B1 and B2. The 54 healthy participants (30 males and 24 females) who contributed the samples measured in B1 have already been described previously.[18] In B2, plasma samples from 128 additional healthy participants (49 males and 79 females) from 42 families were measured. In this cohort, there were 16 monozygotic twin pairs, 26 dizygotic twin pairs, and 44 nontwin siblings. The average age of the twins in the cohort of whom samples were measured in B2 was 18.2 years (standard deviation (SD), 0.2); the average age of the siblings was 19.5 years (SD, 4.8). In B1, for LC−MS analysis two aliquots were taken of the plasma sample from each individual participant, which were then divided into two measurement batches where each batch contained one aliquot of each study sample. In B2, on the other hand, only one aliquot of each study sample was processed and analyzed in one measurement batch. Furthermore, following every other of the QC sample aliquots consisting of B2 study samples, aliquots were inserted of the QC sample that had been measured in B1 as well and that thus consisted of B1 individual study sample aliquots (sample pretreatment was performed for this B1 QC sample in B1 and in B2 separately). This B1 QC sample thus underwent an additional freeze−thaw cycle between B1 and B2. As a measure of experimental error, for each detected lipid compound relative standard deviations (RSDs) were computed for B1 of the IS-corrected measurements in B1 of the pooled QC sample prepared from individual study samples measured in B1, and for B2 of the IS-corrected measurements of the pooled QC sample prepared from samples measured in B2. In B2, for NMR analysis following each of the QC sample aliquots consisting of B2 study samples, samples were inserted of in total 12 participants that had already been analyzed in B1. These samples thus underwent an additional freeze−thaw cycle between B1 and B2.

Equating Data from B1 and B2. Our equating method lets the data for each variable assume the same distribution in all blocks, by averaging the distributions for that variable in all blocks. An algorithm to achieve this has been presented by Bolstad et al.[21,22] This algorithm was based on the principle of the quantile−quantile plot (Q−Q plot). Generally stated, quantiles are the values marking the boundaries between regular intervals of the cumulative distribution of a data sample. That is, when dividing ranked data into a number of subsets, then the quantiles are the values at the boundaries between consecutive subsets. In a Q−Q plot, the quantile values of two distributions are plotted against each other; the number of quantiles plotted equals the number of data points in the smaller data sample (the quantile values in the larger data sample are found by linear interpolation).[23,24] If in the Q−Q plot the points defined by the values of corresponding quantiles in both data samples all lie on a straight diagonal line, then the distributions of both samples are highly similar; if they do not, then the distributions are dissimilar. In the algorithm as presented by Bolstad et al., the averaging of data distributions is achieved by projecting the corresponding quantile values of all distributions onto a scalar multiple of the unit vector (a, possibly multidimensional, analogue of the diagonal in the Q−Q plot) (Figure 1).[21,22] Then, the averaged quantile values are substituted for the original values that are in the subsets belonging to the corresponding quantiles in the data samples under consideration. Thus, the original ranking of the data points in the data samples to be combined is retained. The result is that the distributions of all data samples become equal, or—in the case of different numbers of observations per data sample—almost equal. This algorithm is usually applied in an "omics" context to make the distributions of different objects equal over all measured variables, that is, for "normalization". Examples of this application are found, e.g., in the fields of genomics (normalization of gene probe

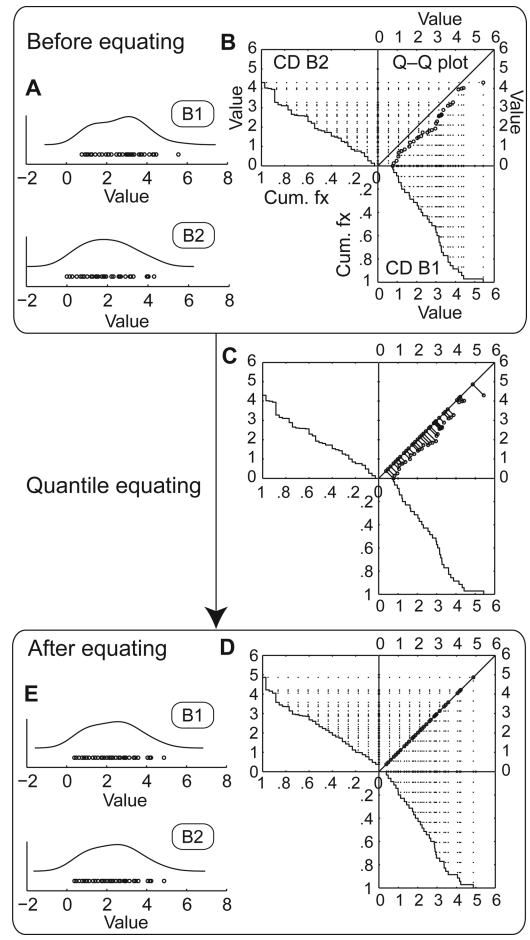

Figure 1. Action of quantile equating algorithm schematically illustrated: Data samples B1 and B2 have different distribution shapes (panel A). The cumulative distributions (CD) corresponding to these distributions are plotted against each other in the quantile−quantile plot (Q−Q plot) in panel B. Quantile equating is attained by projecting the values of corresponding quantiles onto a scalar multiple of the unit vector (the diagonal line in the Q−Q plot) in panel C. Then, the projected (averaged) quantile values are substituted for the original values in the subsets belonging to each quantile. Thereby, the distributions of B1 and B2 become equal, as is illustrated with equal cumulative distributions (panel D) and equal kernel densities (panel E). Data from ref 35. CD, cumulative distribution; Q−Q plot, quantile−quantile plot; Cum. fx, cumulative fraction. The axis labels as in panel B apply to panels C and D as well.

(20) Hendriks, M. M.; Smit, S.; Akkermans, W. L.; Reijmers, T. H.; Eilers, P. H.; Hoefsloot, H. C.; Rubingh, C. M.; de Koster, C. G.; Aerts, J. M.; Smilde, A. K. *Proteomics* **2007**, *7*, 3672–3680.

(21) Bolstad, B. M. Division of Biostatistics, University of California, Berkeley. Probe level quantile normalization of high density oligonucleotide array data. Unpublished work, 2001.

(22) Bolstad, B. M.; Irizarry, R. A.; Astrand, M.; Speed, T. P. *Bioinformatics* **2003**, *19*, 185–193.

(23) Cleveland, W. S. *The Elements of Graphing Data*, 2nd ed.; Hobart Press: Summit, NJ, 1994; pp 133−149.

(24) Wilk, M. B.; Gnanadesikan, R. *Biometrika* **1968**, *55*, 1–17.

intensity distributions between oligomicroarrays, over all gene probes)[22,25−27] and of peptidomics (normalization of peptide intensity distributions between analytical samples, over all detected peptides).[28] However, we introduce the use of this algorithm for equating, that is, for making the distributions of the same variable (NMR feature or lipid) equal over all sets of objects (sets of study samples in all blocks). Because our method is conceptually akin to what is known in psychometrics as "quantile equating" or "equipercentile equating",[16,29] we will refer to it as "quantile equating" as well. Of note, in quantile equating in a psychometrical context the aim is not to make the distributions of the same variable equal for all sets of objects but to provide transformations by which equivalent scores can be found on different versions of the same test. We used the "normalize.quantiles" function, which was written by the first author of the original publications,[21,22] to perform quantile equating. This function is part of the "preprocessCore" package, which is a component of the Bioconductor software suite (version 2.1)[30] running in the statistical environment R (version 2.6.2).[31] For its originally intended purpose, i.e., for *normalization*, the "normalize.quantiles" function is applied simultaneously to all *objects* (study samples). To perform *equating*, however, we applied this function to the *variables*. Moreover, we applied the function to the B1 and B2 data for each variable separately. In case of the LC−MS data, replicate measurements of the individual study samples in B1 were first averaged before equating, whereas in case of the NMR data unaveraged replicates were equated. Data for samples measured in B1 as well as in B2 (for example, QC samples prepared on basis of pooled aliquots of B1 individual study samples) were omitted from all B2 data sets before equating for the following reason. If the composition of QC samples changes differently between measurement blocks with respect to the composition of individual study samples, then QC samples are not representative for the samples measured in all blocks. In this paper, we show an example of this in case of plasma NMR spectroscopy, where repeatedly measured samples underwent an additional freeze−thaw cycle between B1 and B2 with respect to the individual samples measured in B2. If we would have left the data for these repeatedly measured samples in the B2 block, these data would have influenced the B2 data distributions and thereby would have distorted the result of quantile equating. We did not remove the B1 and B2 measurement data for the QC samples prepared on basis of samples measured in each block, because these helped to visualize the beneficial effects of quantile equating in making combinable B1 and B2 data sets.

Evaluation of Comparability of Data Sets. The comparability of data sets obtained with the same analytical method but in different measurement blocks was evaluated using various methods. At the univariate level, before quantile equating we assessed to which extent the relationship between data distributions of both measurement blocks was nonlinear using the Pearson correlations between the ranked quantile values of both measurement blocks. Due to the nature of quantile equating, after equating the correlations between the B1 and B2 quantile values are always equal to 1. We characterized the extent to which nonlinear relationships between the distributions as well as other differences between the data from both measurement blocks before equating gave rise to differences at the multivariate level, using a strategy proposed by Jouan-Rimbaud et al.[32] In this strategy, data sets are compared in the principal component (PC) space using three continuous parameters that each can take a value between 0 and 1, where a zero value indicates low similarity of the evaluated data sets and a value of 1 suggests perfect similarity. The first parameter ("*P*") is based upon the comparison of principal components analysis (PCA) loadings patterns, the second parameter ("*C*") is based upon the comparison of variance−covariance matrices, and the third parameter ("*R*") characterizes the similarity in location of the centroids of the data sets. The degree of success of quantile equating in making data from both measurement blocks comparable, was characterized using these multivariate parameters as well. We used a 2% increase in total variance explained by the model as a criterion to estimate the number of PCs for which these parameters were to be computed (PLS_Toolbox version 3.5, Eigenvector Research, Wenatchee, WA). Furthermore, the success of the equating procedure was visualized by the results of PCA on the combined (concatenated with the variables as the shared mode) data sets originating from different measurement blocks. For this PCA, replicate measurements were averaged. LC−MS data were then mean-centered, whereas NMR data were autoscaled. These different types of scaling were applied to the respective types of data because this enhanced the visibility of the between-block effects prior to equating. All PCA were carried out using the PLS_Toolbox for MATLAB (version R2006b, The Mathworks, Natick, MA).

RESULTS AND DISCUSSION

Analytical Data. In ref 18, the data denoted in the current paper as the B1 LC−MS data have already been presented. The 61 different lipids that were detected in the chromatograms in B1 were detected in B2 as well.[18] Lipids from the following classes were detected: lysophosphatidylcholines (LPC), phosphatidylcholines (PhC), sphingomyelins (SPM), cholesterol esters, and triglycerides (TG). Throughout the manuscript, lipids are denoted as follows: the number of carbon atoms as well as the number of double bonds in the fatty acid, separated by a colon (e.g., C36:5) is followed by the class abbreviation (e.g., PhC).[13] The data for C16:0_LPC and C52:2_TG were excluded from further analysis because their responses displayed a systematic trend in the QC sample measurements in B2, resulting in high RSDs. In B1, the

(25) Cawley, S.; Bekiranov, S.; Ng, H. H.; Kapranov, P.; Sekinger, E. A.; Kampa, D.; Piccolboni, A.; Sementchenko, V.; Cheng, J.; Williams, A. J.; Wheeler, R.; Wong, B.; Drenkow, J.; Yamanaka, M.; Patel, S.; Brubaker, S.; Tammana, H.; Helt, G.; Struhl, K.; Gingeras, T. R. *Cell* **2004**, *116*, 499–509.

(26) Kim, J. W.; Tchernyshyov, I.; Semenza, G. L.; Dang, C. V. *Cell Metab.* **2006**, *3*, 177–185.

(27) Chen, H. Y.; Yu, S. L.; Chen, C. H.; Chang, G. C.; Chen, C. Y.; Yuan, A.; Cheng, C. L.; Wang, C. H.; Terng, H. J.; Kao, S. F.; Chan, W. K.; Li, H. N.; Liu, C. C.; Singh, S.; Chen, W. J.; Chen, J. J.; Yang, P. C. *N. Engl. J. Med.* **2007**, *356*, 11–20.

(28) Higgs, R. E.; Knierman, M. D.; Gelfanova, V.; Butler, J. P.; Hale, J. E. *J. Proteome Res.* **2005**, *4*, 1442–1450.

(29) Angoff, W. H. In *Educational Measurement*, 2nd ed.; Thorndike, R. L., Ed.; American Council on Education: Washington, DC, 1971; pp 562−600.

(30) Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; Hornik, K.; Hothorn, T.; Huber, W.; Iacus, S.; Irizarry, R.; Leisch, F.; Li, C.; Maechler, M.; Rossini, A. J.; Sawitzki, G.; Smith, C.; Smyth, G.; Tierney, L.; Yang, J. Y.; Zhang, J. *Genome Biol.* **2004**, *5*, R80.

(31) R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2008.

(32) Jouan-Rimbaud, D.; Massart, D. L.; Saby, C. A.; Puel, C. *Chemom. Intell. Lab. Syst.* **1998**, *40*, 129–144.

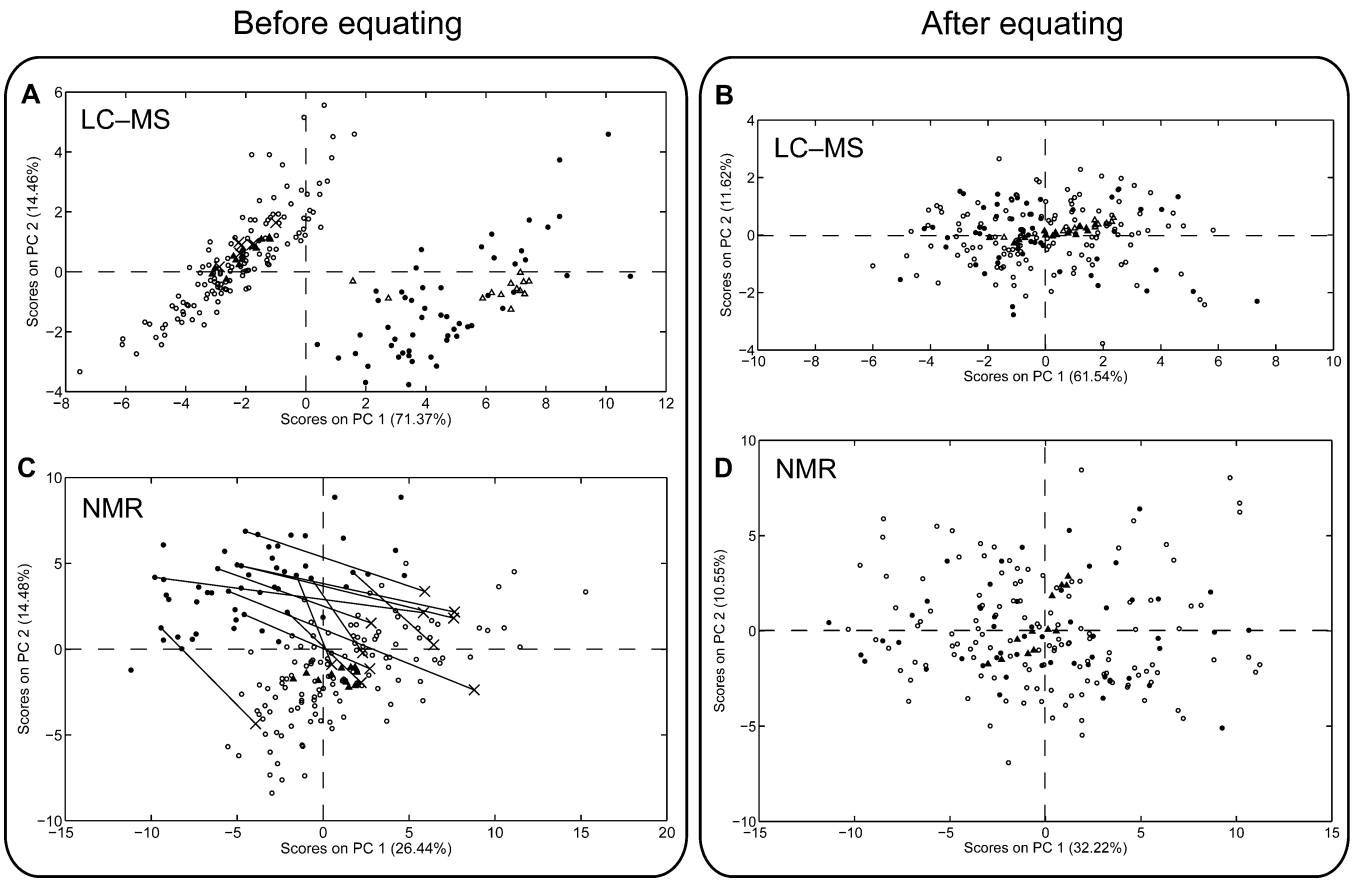

Figure 2. PCA scores on PC1 and PC2 for the combined (concatenated) B1−B2 data sets before (panels A and C) and after (panels B and D) quantile equating. Panels A and B, plasma LC−MS data; panels C and D, plasma NMR data. In panels A and B, B1 QC sample aliquots measured in B1 are indicated by (△). In panel C, scores based on NMR measurements of individual plasma samples that were measured in both B1 and B2 are connected by lines. The percentages of variance explained by the respective PCs are given between brackets in the axes labels. PC1−PC2 loadings plots are given in the Supporting Information (Figures S4 and S5). ●, B1 individual study sample; ○, B2 individual study sample; ▲, B2 QC sample aliquot measured in B2; ×, B1 QC sample aliquot (panel A) or B1 individual study sample (panel C) measured in B2.

mean RSDs for the remaining 59 lipids as computed on basis of the measurements of the QC sample prepared in B1 were 13.3% (SD, 5.6; range, 5.2−25.5%). Notably, the RSDs of all LPCs, PhCs, and SPMs were below 15%. In B2, the mean RSDs of these same 59 lipids, computed on basis of the measurements of the QC sample prepared in B2, were 7.5% (SD, 1.4; range, 4.9−10.9%).

In the plasma NMR data, after application of the "80% rule", 75 features (variables) were kept for analysis.

B1−B2 Comparison before Equating. *PCA Scores Plots.* Panels A and C of Figure 2 display the PCA scores plots for the LC−MS and the NMR plasma data, respectively, before equating. As expected, the scores of almost all pooled B1 and B2 QC sample aliquots are in the centers of the clusters corresponding to B1 and B2, respectively. However, in particular in case of the LC−MS data, the scores of the measurements from both blocks display notable separation along the PC1 axis (Figure 2A). This phenomenon might have been caused, for example, by slightly different IS concentrations. Another possible cause is that for each block a separate target table was constructed on basis of the QC sample measurements in that block. This might have led to different detection thresholds for the same peaks in both blocks and thereby to systematic differences in peak integrals. The scores based on the B1 and on the B2 plasma NMR measurements overlapped only partially (Figure 2C). This may have been caused, at least in part, by different CPMG parameter sets in both blocks. Furthermore, in Figure 2C, it can be observed that the NMR measurements in B2 of the 12 individual samples that were measured in B1 as well are not representative for the measurements in B2. We suspect that this is among others due to the additional freeze−thaw cycle that these repeatedly measured samples underwent and that is known to affect plasma NMR spectra.[33] Therefore, Figure 2C gives a visual illustration of a case where methods that employ such repeatedly measured samples for equating, e.g., the method described in ref 14, cannot be used.

B1−B2 Correlation of Quantile Values. The average Pearson correlation for all variables between the B1 and the B2 quantile values before equating was 0.97 (SD, 0.03) for the LC−MS data and 0.92 (SD, 0.09) for the plasma NMR data. In case of the LC−MS data, notably a group of TGs displayed nonlinear relationships between the quantile values of both blocks (Supporting Information Table S3A). Among the lipids, TGs are particularly likely to display nonlinear differences in data distribution shapes among data blocks because they can form dimers during ionization and MS detection. This effect is dependent on

(33) Deprez, S.; Sweatman, B. C.; Connor, S. C.; Haselden, J. N.; Waterfield, C. J. *J. Pharm. Biomed. Anal.* **2002**, *30*, 1297–1310.

Table 1. B1−B2 Similarity of Data Sets in PC Space before and after Quantile Equating[a]

A (LC−MS Data, before Equating)

	1 PC	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs
P	0.9615	0.9423	0.9339	0.9315	0.9463	0.9513
C	0.9829	0.9504	0.6682	0.6527	0.6181	0.4553
R	0	0	0	0	0	0

B (LC−MS Data, after Equating)

	1 PC	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs
P	0.9958	0.9952	0.9897	0.9926	0.9954	0.9941
C	0.9984	0.9935	0.9902	0.9844	0.9645	0.9392
R	0.9997	0.9985	0.9988	0.9988	0.999	0.9988

C (^{1}H NMR Data, before Equating)

	1 PC	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs	7 PCs
P	0.949	0.9143	0.9125	0.9057	0.8919	0.8962	0.8936
C	0.9964	0.9947	0.713	0.6732	0.5372	0.3266	0.2944
R	0	0	0	0	0	0	0

D (^{1}H NMR Data, after Equating)

	1 PC	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs	7 PCs
P	0.9892	0.951	0.975	0.97	0.9684	0.9684	0.9679
C	0.999	0.9716	0.805	0.721	0.6402	0.5964	0.5572
R	0.9996	0.9985	0.9866	0.9857	0.9874	0.9879	0.9881

[a] Sections A and B, similarity of B1 and B2 plasma LC−MS data sets before (section A) and after (section B) quantile equating. Sections C and D, similarity of B1 and B2 plasma ^{1}H NMR data sets before (section C) and after (section D) quantile equating. *P*, B1−B2 similarity of PCA loadings patterns; *C*, B1−B2 similarity of variance−covariance matrices; *R*, B1−B2 similarity of data set centroid locations.

concentration and on ion source tuning. Unlike LC−MS systems, NMR spectrometers are regarded to be linear detectors,[34] implying that signal intensity should be linearly related to compound concentration over the complete dynamic range. Therefore, in case of the NMR data, nonlinear relationships between the distributions of the B1 and the B2 data at lower intensities (Supporting Information Table S3B) might have been caused by differences in the sensitivity of the NMR probe heads used for the acquisitions of the NMR data between both blocks, as well as by differences in peak detection thresholds between both blocks.

Multivariate Parameters. The values of parameters that characterize the similarity of the B1 and B2 data sets in the PC space before and after quantile equating are given in Table 1. For both the LC−MS data and the plasma NMR data, the values for the *P* parameter as well as the values for the *C* parameter with inclusion of two PCs suggest that the structures of the B1 and B2 data are already comparable before equating (Table 1, sections A and C). This is important because it suggests that the compositions of the object groups are indeed balanced between both measurement blocks. Therefore it might be reasonable to assume that with application of the quantile equating method, relatively much analytical between-block variation will be removed with respect to biological variation. However, the zero values for the *R* parameter in case of both the LC−MS as well as the NMR data suggest that there is a multiplicative difference between the B1 and B2 data, which is in concordance with what can be observed in the PCA scores plots on the combined data sets (Figure 2, panels A and C). Moreover, in Table 1, sections A and C, the values for the *C* parameter decrease considerably with inclusion of more than two PCs, suggesting that the higher PCs are influenced by differences in data distribution shapes between B1 and B2.

B1−B2 Comparison after Equating. *PCA Scores Plots.* After quantile equating of the data, the systematic nonbiological differences between the B1 and B2 data are not manifest anymore in the PCA scores plots (Figure 2, panels B and D). In these plots, the scores based on the individual study samples measured in B1 and B2 are dispersed among each other. Also, the scores based on the measurements of the pooled QC samples in both B1 and B2 are located in the centers of the plots. This is consistent with the expectation that the B1 and B2 pooled QC samples should represent the average sample measured in each of the blocks. Given that this expectation is correct, the location in the centers of the plots of the QC sample measurement scores from both B1 and B2 in turn is a direct consequence of making the data distributions of each variable equal for both blocks by quantile equating.

Multivariate Parameters. For both LC−MS and NMR, the increase in the values of the *R* parameter after equating (Table 1 sections B and D) suggests that in particular the distance between the centroids of the B1 and B2 data sets has decreased. The values for the *P* and *C* parameters have increased as well. The values for all parameters are not equal to 1 after equating, which is consistent with the notion that although our univariate equating method causes equal or nearly equal data distributions among data blocks at the univariate level, the ranking of objects at this univariate level is retained. Therefore, differences among data blocks at the multivariate level are not necessarily removed by univariate quantile equating as well.

CONCLUSIONS

Combination of semiquantitative metabolomics data sets originating from different measurement blocks where the same metabolites have been measured can be challenging due to nonbiological systematic differences among the blocks. These differences are caused by unwanted, though sometimes practically unavoidable, between-block differences in experimental conditions. We have presented a solution for such data combination problems in the form of the quantile equating method. We have demonstrated the successful application of the quantile equating method to LC−MS and ^{1}H NMR metabolomics data obtained in human plasma samples. We successfully applied our equating method to urine ^{1}H NMR metabolomics data as well (see the Supporting Information for methods and results). It is conceivable that the quantile equating method is equally applicable for other types of semiquantitative metabolomics data, e.g., GC/MS data. Due to its univariate nature, this equating method will remain to provide satisfactory results even when the data sets to be combined contain data for (much) larger numbers of variables than the examples considered in this article. Moreover, the applicability of the equating method presented in this article may not be limited to data from metabolomics studies. For example,

(34) Mehr, K.; John, B.; Russell, D.; Avizonis, D. *Anal. Chem.* **2008**, *80*, 8320–8323.
(35) Frisby, J. P.; Clatworthy, J. L. *Perception* **1975**, *4*, 173–178.

in DNA methylation measurements in the context of epigenetics studies the data distributions may vary between arrays and equating methods have the potential to correct the data obtained in such experiments. Of course, the possibility to apply equating methods in an "omics" context leaves unimpeded the importance of good analytical practice. This includes that, if possible, all study samples should be measured in one block to minimize process variability. However, in a typical large metabolomics study, where in total hundreds or thousands of samples are measured, it is often not feasible both from a practical and cost perspective to measure new and previously measured samples together in one block. Because of such practical limitations, and because not all systematic differences between measurements in different analytical blocks can be prevented by good analytical practice alone, we believe that equating methods have the potential to enable joint analysis of valuable data sets, which would not be possible without using such methods.

ACKNOWLEDGMENT

We thank all the twins and siblings who participated in this study. We acknowledge support from The Netherlands Bioinformatics Centre (NBIC) through its research programme BioRange (project no. SP 3.3.1), Spinozapremie NWO/SPI 56-464-14192, the Center for Medical Systems Biology (CMSB), Twin-family database for behavior genetics and genomics studies (NWO-MaGW 480-04-004), and NWO-MaGW Vervangingsstudie (NWO no. 400-05-717).

SUPPORTING INFORMATION AVAILABLE

Methods and results for application of the quantile equating method to data from ^{1}H NMR analysis of urine samples; for plasma LC−MS and ^{1}H NMR data, PC1−PC2 loadings plots and lists with variables having lowest B1−B2 correlation of quantile values before equating; typical examples of plasma ^{1}H NMR spectra. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review October 16, 2009. Accepted December 3, 2009.

AC902346A