Combination of metabolomics datasets from different measurement series

H Draisma[1], F van der Kloet[1,2], T Reijmers[1], I Bobeldijk-Pastorova[2], J Meulman[3], F Estourgie–van Burk[4,5], J van der Greef[1], D Boomsma[5], E Spies–Faber[2], J Vogels[2], T Hankemeier[1]

[1] Leiden University, LACDR, Leiden, The Netherlands; [2] TNO Quality of Life, Zeist, The Netherlands; [3] Leiden University, Mathematical Institute, Leiden, The Netherlands;
[4] Department of Paediatric Endocrinology, Institute for Clinical and Experimental Neurosciences, VU University Medical Center, Amsterdam, The Netherlands;
[5] Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands

Contact: **h.draisma@lacdr.leidenuniv.nl**

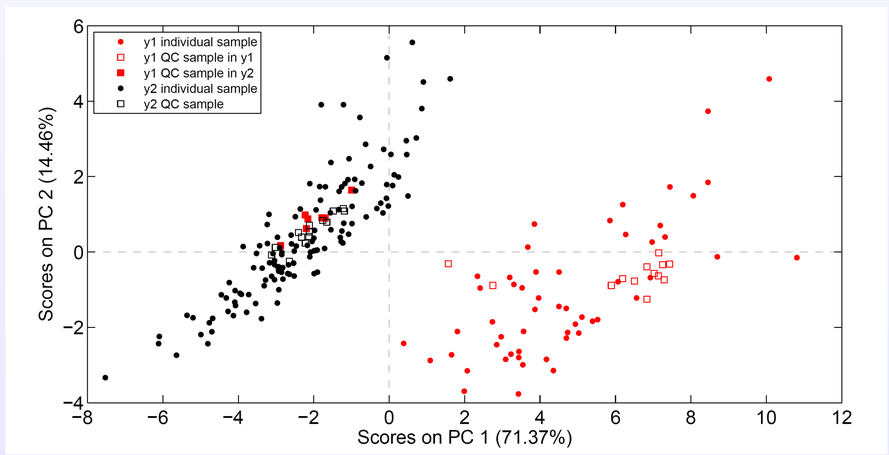
Introduction

Combination of data from different sources is an important topic in systems biology. To increase the *power* of statistical analyses, combination of datasets from different measurements on different groups of objects (patients, samples,…) is desired.

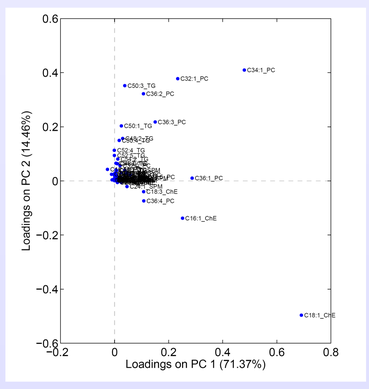
It is often not possible to guarantee comparability of such datasets because of the impossibility to make full calibration models [Sangster *et al*, The Analyst 2006: 1075–1078].

Indeed, the impossibility to do calibration model transfer can for example lead to serious between-batch effects **(Fig. 1A)**. As an alternative, we propose the method of quantile *equating* [Van der Linden, Psychometrika 2000: 437–456] to make datasets comparable that originate from different measurement series on similar groups of objects. We illustrate this method using data from two batches ('y1' and 'y2') of human blood plasma samples that were analyzed with almost one year in between by the same liquid chromatography–mass spectrometry (LC–MS) platform.

1A



1B



1C

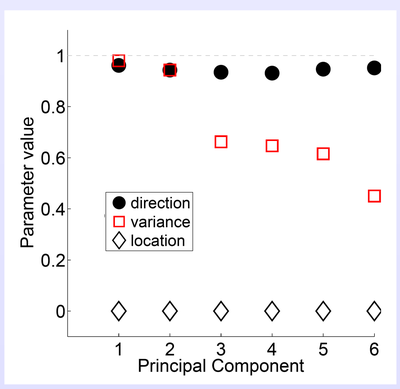


Figure 1. Results of multivariate analysis of the data from y1 and y2 before equating.
A: PCA scores. **B**: PCA loadings. **C**: Values of parameters indicating: similarity of loadings patterns ('direction'); similarity of var-covar matrices [Box's *M*] ('variance'); and Mahalanobis distance between datasets ('location'). All parameters: 0 = 'dissimilar', 1 = 'similar'.
PCA results are based on relative responses in y1 and y2 (data meancentered). Abbr.: LPC, lysophosphatidylcholines; PC, phosphatidylcholines; SPM, sphingomyelins; ChE, cholesterol esters; TG, triglycerides.

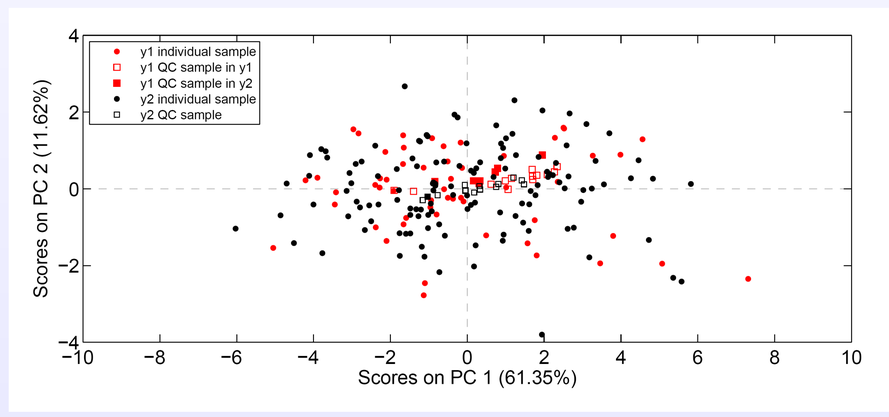
Lipid LC–MS analysis was performed on the samples of healthy volunteers (Netherlands Twin Register) (*N*=54 in y1 and *N*=128 in y2) [Draisma *et al*, OMICS 2008: 17–31]. Data on 59 different lipids common for both years were corrected using class-specific internal standards (5 lipid classes; in total 4 IS used).

We evaluated the success of equating using three parameters for the similarity of datasets in the multivariate space [Jouan-Rimbaud *et al*, Chemom Intell Lab Syst 1998: 129–144] **(Figs. 1C & 2C)**.

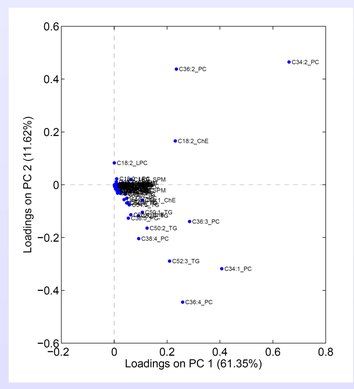
The values for these parameters prior to equating suggested that the object groups were similar, but that in particular their locations differed.

After equating, notably the Mahalanobis distance between the y1 and y2 data had decreased dramatically, indicating that the offset as seen in **Figure 1A** had been corrected for.

2A



2B



2C

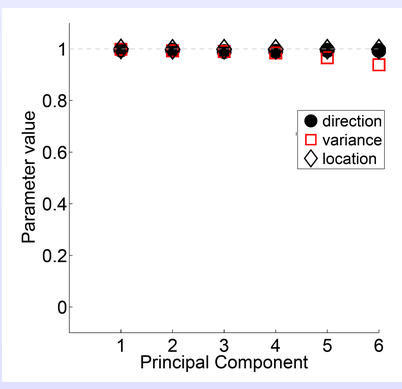


Figure 2. Results of multivariate analysis of the data from y1 and y2 after quantile equating. For explanation, see the legend to Figure 1.

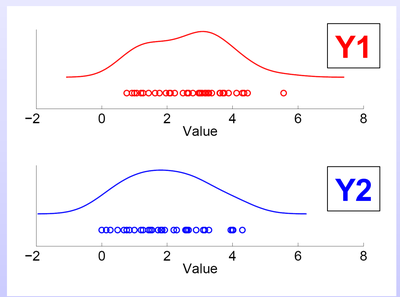
Acknowledgment

We would like to thank all twins and their siblings who participated in this study.

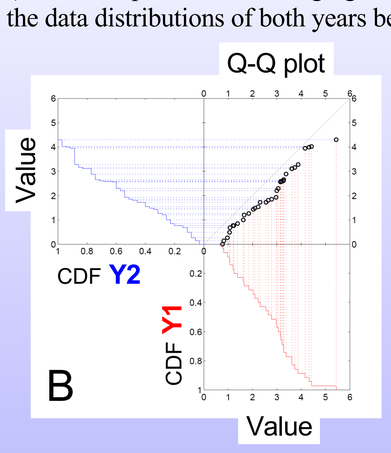
Univariate quantile equating

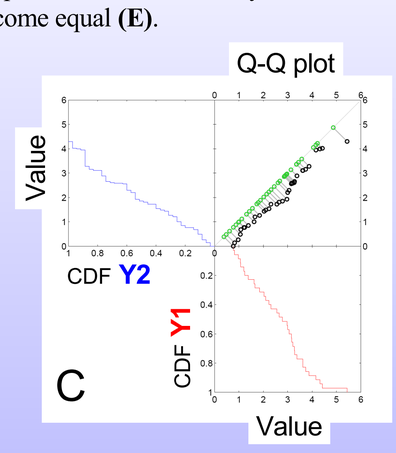
The data were equated univariately by quantile equating, based on the quantile normalization method developed by Bolstad *et al* [Bolstad *et al*, Bioinformatics 2003: 185–193]. The corresponding quantile values of both years were projected onto the unit vector **(C)**. This is equivalent to averaging the quantile values for both years. The averaged value of each quantile was then substituted for each of the individual data values belonging to that quantile **(D)**. The result is that the data distributions of both years become equal **(E)**.

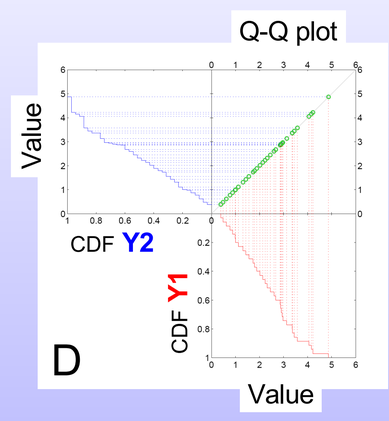
A

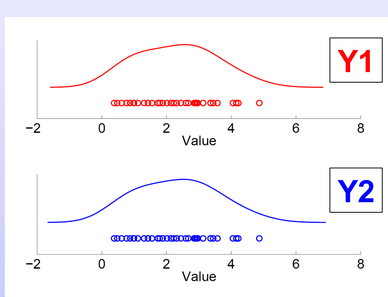


Scatter plots and kernel density plots for Y1 and Y2 data prior to equating (data from [Frisby & Clatworthy, Perception 1975: 173–178]).









E

Leiden / Amsterdam Center for Drug Research
P.O. Box 9502, 2300 RA Leiden
The Netherlands
Universiteit Leiden
Vrije Universiteit Amsterdam

Netherlands Metabolomics Centre

CENTRE FOR Medical Systems Biology

Universiteit Leiden

nbic netherlands bioinformatics centre

BioRange activity of NBIC

This work was supported by the Netherlands Bioinformatics Centre (BioRange project SP 3.3.1)