High-throughput $^1$H NMR-based metabolic analysis of human serum and urine for large-scale epidemiological studies: validation study

Richard H Barton, 1* Jeremy K Nicholson, 1 Paul Elliott 2 and Elaine Holmes 1

Accepted 10 December 2007

Background Metabolic profiling of biofluid specimens is an established method for investigating disease states in clinical studies but is only recently being applied to large-scale human population studies. As part of protocol development for the UK Biobank study, a $^1$H nuclear magnetic resonance (NMR)-based metabonomic analysis of specimen storage effects and analytical reproducibility was carried out using urine and serum specimens from 40 volunteers.

Methods Aliquots of each specimen were stored for $t=0$ and $t=24$ h at 4°C prior to freezing, and in the case of serum samples for a further 12 h ($t=36$), to determine whether the storage times affected specimen composition and quality. A blinded split-specimen matching exercise was implemented to assign candidate spectral pairs stored for different times using multivariate statistical analysis of the NMR data.

Results Using a chemometric strategy, split specimens at time $t=0$ and $t=24$ or 36 h after storage at 4°C were easily paired and the split-specimen matching task was reduced to a workable size. $^1$H NMR profiling established that the $t=24$ h urine and serum groups showed no systematic metabolite changes, indicating biochemical stability. Some small differences in serum specimens stored for $t=36$ h at 4°C were detectable only by multivariate analysis, and were attributed to generalized alterations in proteins and protein fragments, and possibly trimethylamine-N-oxide. No other specific metabolite was implicated.

Conclusions For the purposes of NMR-based analysis, storage of urine and serum for up to $t=24$ h at 4°C does not detectably affect the metabolic profile and the methodology is robust. Future application of multivariate methods to data-rich studies should substantially enhance information recovery from epidemiological studies.

Keywords Metabonomics, urine, serum, NMR, molecular epidemiology, multivariate

1 Biomolecular Medicine, SORA Division, Faculty of Medicine, Imperial College London, Exhibition Road, London SW7 2AZ, UK.

2 Department of Epidemiology and Public Health, Faculty of Medicine, Imperial College London, St Mary’s campus, Norfolk Place, London W2 1PG, UK.

* Corresponding author. Biomolecular Medicine, Faculty of Medicine, Imperial College London, Exhibition Rd, London SW7 2AZ. E-mail: r.barton@ic.ac.uk
Introduction

Molecular epidemiology aims to enrich the data available in epidemiological investigation by incorporating detailed information on the metabolic status of an individual. Amongst the challenges facing modern epidemiological studies are issues of large sample sizes, high throughput and high specimen numbers, multiple field sites and investigators and practical aspects such as handling and stability of specimens in storage, transit of specimens to and between laboratories and analytical variability. Attention needs to be paid to specimen stability during long analysis periods or where there is long-term specimen archiving. Particular challenges face longitudinal studies as analytical technologies may vary over time, and features such as instrumental drift and variation in detection limits can affect analytical results and sensitivity. This source of variation needs to be differentiated from the biological variation both within and between individuals that may underlie differences in disease risk.

Metabolic profiling has been used in animal and human-based studies of health and disease, and has been shown to be a valuable approach for characterizing disease and generating candidate biomarkers of physiological or pathological state.1–3 Such studies typically involve some form of high-resolution analytical spectroscopy, generally Nuclear Magnetic Resonance (NMR) spectroscopy or Mass Spectrometry (MS), in combination with multivariate mathematical modelling. This article focuses on NMR-based metabolic profiling of specimen storage effects done as part of the UK Biobank pilot study conducted on urine and blood serum specimens obtained from 40 volunteers.

Methods

Aliquots of each specimen were analysed at time $t=0$, or $t=24\text{h}$ after storage at $4^\circ\text{C}$, prior to freezing at $-80^\circ\text{C}$, to determine whether a delay in freezing up to 24h produced any systematic compositional changes.4 The study was subsequently extended to examine 36 serum specimens stored at $4^\circ\text{C}$ for $t=36\text{h}$ prior to freezing.

Specimen collection and preparation

Specimens were provided for the Biobank pilot study by Dxs Ltd (UK), and shipped to Imperial College London on dry ice. Specimens were then stored at $-85^\circ\text{C}$ prior to preparation for NMR acquisition. Serum and urine specimen preparation was similar to that of prior studies,3,5 except that in this case all specimens were prepared directly into 5 mm Norell HP507 NMR tubes.

NMR spectroscopic analysis and data analysis

All spectra were acquired at 600.289 MHz using a TXI probe on a Bruker DRX600 spectrometer, using the Bruker ‘noesypr1d’ pulse programme. Acquisition parameters and NMR processing, including spectral referencing, were implemented similarly to prior studies,3,5 as was subsequent data processing. Water suppression was carried out by irradiation during the recovery delay time and the first increment of the noesy pulse sequence mixing time. Primary acquisition parameters included 64 scans over a 20ppm sweep width into 32k data points, with a recovery delay of 2s. The free induction decay (FID) was zero-filled once and Fourier transformed with a line broadening of 1 Hz for plasma and 0.3 Hz for urine. All resultant 1D NMR spectra were corrected for phase and baseline distortions using in-house Matlab routines (Dr T. Ebbels and Dr H. Keun). Urine spectra were further automatically referenced to TSP at 0.0ppm., whereas the serum spectra were hand referenced to the glucose anomeric doublet at 5.23 ppm. To verify reproducibility of the analytical platform, random samples were also physically re-presented to the spectrometer for a repeat acquisition. Data were either reduced into integrated regions of equal width (0.04 ppm. for urine or 0.01 ppm. for serum) using AMIX (Analysis of MIXtures, ver. 3.1, Bruker Biospin Germany), or imported at full resolution into Matlab using in-house routines (Dr O. Cloarec). No variance re-scaling was applied for the analyses described, and most analysis was carried out on normalized data. However, raw-data spectral comparisons were also carried out using Xwinnmr for some specimens, as discussed subsequently. Data analysis methods included Principal Components Analysis (PCA) and Projection to Latent Structure Discriminant Analysis (PLS-DA), as well as Orthogonally Filtered PLS-DA (O-PLS-DA).6,7 PCA modelling is unsupervised and maximizes the explained variance of data for each successive component of the model, removing this variation at each step and continuing the analysis of the residual variance. PLS-DA methods use knowledge of specimen class or origin to define boundaries between two or more classes of observation (in this case specimens held for two different lengths of time prior to freezing). Each spectrum is displayed as a single coordinate in the scores plot, and the proximity of any two samples is directly related to their similarity in spectral profile and hence biochemical composition. The coefficients plots generated by this method also provide information on the nature of the differences between classes, and in this case will identify metabolites that alter with storage time, providing an indication of time-related biochemical degradation of the specimen.

An initial data reduction step was applied prior to PCA and PLS-DA modelling where spectra were integrated (bucketed) over defined spectral areas to search for any systematic differences in metabolic profiles. To refine integrated data models, if a PLS-DA model of integrated data exhibited any
systematic separation between specimen storage groups, then O-PLS-DA modelling at full spectral resolution was carried out in the Matlab 7.0 environment (Mathworks, Natick, MA) using proprietary algorithms (Dr Ebbels and Dr Cloarec) to verify the chemical basis of the observed discrimination at the highest possible resolution.6–8

Split specimen pairing

It was anticipated that inter-person variation would be greater than variation introduced by length of storage time within the first 24 h, therefore the first part of the study was implemented as a blinded split-specimen exercise between the specimens frozen at \( t = 0 \) h (40 urine; 40 serum) and their matched aliquots held for \( t = 24 \) h at 4°C prior to freezing. It was anticipated that this exercise might be more difficult for serum than for the urine data, due to the potential for variations in wetting, homogeneity and minor freeze–thaw-induced macrostructure degeneration in sera affecting sampling. A similar exercise was subsequently implemented for a further 36 serum specimens held at 4°C for \( t = 36 \) h, matched against the \( t = 0 \) h specimens. PCA was chosen as a first approach, since it supplies information about both the ‘tightness’ of pairings between the split specimens (i.e. the general similarity of the specimen profiles between the two conditions), and also gives an indication if there is other significant intrinsic structure present in the data relating to storage time. Specimen class information was not used in the calculation of the PCA models. We trialled but did not employ hierarchical clustering analysis as this may give differing results depending on the cluster algorithm scaling used.

In order to rapidly implement the identification of split specimen pairs, an iterative application of 2-component PCA models on integrated data was employed, with the split groups for \( t = 0 \) or \( t = 24 \) h storage prior to freezing. The strategy was implemented by assigning all unambiguous class pairs starting from the outside regions of the PCA model clusters, and moving inwards until no further unambiguous pairs could be identified. All assigned pairs were then documented and excluded and the residual specimens remodelled using PCA. The pairing procedure and 2-component PCA remodellings were repeated until no further clear pairings were achieved. This provided an efficient strategy in contrast to an exhaustive reducing pairwise comparison, which would require a large number of pairwise comparisons to be carried out; e.g. for 40 pairs of specimens a complete search yields \( (40 + 39 + \cdots + 2) = 819 \) required comparisons. The specimen residual unmodelled variance (DModX measure, SIMCA) for each specimen was used to indicate the most likely pairings in the remaining specimens and final confirmation of these was carried out by direct pairwise spectral comparisons within Xwinnmr. The overall pairing procedure applied to both the urine and serum datasets is summarized in the flowchart in Figure 1.

Analysis of treatment group differences

In order to enhance the identification of potential differences in biochemical composition caused by delayed freezing after specimen collection, PLS-DA modelling was then employed, using knowledge of the specimen classification. Discriminant analysis maximizes any separation present between defined membership classes (Y matrix) by appropriate weighting of the NMR variables used as the X matrix.9 In this case, we chose to analyse for systematic differences between specimens which were immediately frozen (\( t = 0 \)) or stored e.g. for \( t = 24 \) h at 4°C prior to freezing. The coefficients of the goodness of fit \( (R^2) \) and prediction \( (Q^2) \) were used to assess the model quality i.e. to establish whether any systematic differences were present due to specimen storage and the loadings were used to define which metabolites, if any, changed as a result of the 24 h storage time before freezing.

Results

Representative standard 1D \(^1\text{H}\) NMR spectra for a urine and a serum specimen indicate the signal complexity of the NMR spectral information obtained from each specimen (Figure 2).

Overview of modelling and pairing procedures

The PCA-based pairing procedure for assigning the \( t = 0 \) and \( t = 24 \) h pairs proved a successful strategy. The use of the specimen residual unmodelled variance measure (DModX in SIMCA) as an additional decision parameter was necessary only for the last few pair matchings of serum specimens. Verification was made by direct spectral comparison for the matching of these last serum specimen pairs.

Urine specimens: \( t = 0 \) vs \( t = 24 \) h at 4°C

Identification of the 40 urine split pairs required three successive 2-component PCA models. Full spectral data comparison was employed for a few near-ambiguous model placements. This ease of pairing was reflected in the PCA structure for the full dataset, Figure 3a, where the replicate pairs are well clustered and resolved. This is a result of two main factors: (i) there is high reproducibility in sampling, both for initial aliquoting by micropipette and for post-thaw sampling of urine, and (ii) the metabolic and concentration dynamic ranges of urinary metabolites require a principal component scale which effectively
compresses replicates with already comparatively small variation closer together in the global model space.

Investigation of the PCA model loadings space for the urine data provided a description of the metabolic variation in the urine specimens. The global urine PCA model was dominated by extremes of variation in three main metabolite signal groupings, namely creatinine, hippurate and acetaminophen (paracetamol) glucuronide and sulphate conjugates. There was a secondary and more idiosyncratic scatter of second rank metabolites, and some participants also excreted ethanol, with one volunteer likely to have been drinking rice wine or sake, known to contain a high content of metabolism-resistant ethyl glucoside. Spectral regions for two urine specimens containing respectively ethyl glucoside and two of the primary acetamenophen conjugates are shown in Figure 3c and d.

The PLS-DA model generated for the entire bucketed urine dataset was not able to differentiate the spectra on the basis of storage time (Figure 4a), confirmed by a low level of cumulative explained $Y$ variance, $R^2(Y)$, and a zero predictive ability, with $Q^2(Y)$ becoming increasingly negative for the first four components of the PLS-DA model. Hence whilst small differences between specimen split samples from a single volunteer (i.e. not being exactly co-mapped in the PCA) might be resolvable, there was no indication of any collective (systematic) effect between the two treatment protocol groups, i.e. at $t=0$ and $t=24$ h.

**Serum specimens: $t=0$ vs $t=24$ h at 4°C**

As with the urine data, identification of the split pairs at $t=0$ and $t=24$ h was successful. Following the split specimen pairing, the same type of PLS-DA analysis was carried out for the $t=0$ and $t=24$ h serum specimens as for the urine samples, using both bucketed and full resolution spectra to search for any systematic metabolite differences resulting from the differing storage conditions. There was no indication of any collective structure (Figure 4b) and hence of any systematic determinable differences between the serum specimens under the two storage conditions, as evidenced by the lack of clustering. Again, the pattern of the $Q^2(Y)$ results for the model of Figure 4b confirms a similar inability to predict group membership for storage within the first 24 h.
Figure 2 Examples of standard 1D $^1$H NMR spectra of serum and urine, showing representative structural complexity produced by multiple metabolite signals. The region about the water peak has been excised from the spectra. The serum spectrum even at low shift (0.5–1.5 ppm.) has complex information present as a series of overlapped broad resonances produced by proteins and lipoproteins, as well as superimposed narrow linewidth resonances from low MW species. The lower amplitude high-shift region of each spectrum has been scaled up by a factor of 4 for visual clarity.

Figure 3 Urine split-specimen integrated data PCA model. (a) PCA scores for all $t = 0$ and $t = 24$ h data, coloured by class: black square, $t = 0$ h storage at 4°C; red circle, $t = 24$ h storage at 4°C; (b) loadings for the PCA plot, indicating strongly influential metabolite groupings; (c) NMR spectrum of selected signals from ethyl glucoside, present in a single volunteer urine specimen; (d) NMR spectrum showing selected signals from acetaminophen metabolites in another urine specimen: g = glucuronide, s = sulphate.
Serum specimens: $t=0$ vs $t=36$ h at $4^\circ$C

The split pairing exercise was carried out using three successive PCA models. Again DModX was used to suggest the final-round choices for spectral pair comparisons. A single logging error was detected among the specimens which had been held for $t=36$ h at $4^\circ$C prior to freezing. In this case two unmatched split samples were indicated, and since the $t=0$ vs $t=24$ h split-specimen comparison had shown no such anomaly, it was assumed that the unmatched specimen in the $36$ h cohort was incorrectly labelled, as subsequently verified by introducing full specimen identifications into the model.

PCA co-modelling of all the data for $t=0$, $t=24$ and $t=36$ h specimens also provided evidence that by $t=36$ h, specimen alteration was beginning to occur. A generally increased intra-split scatter of the $t=36$ h specimens was evident compared with that of the $t=24$ h specimens, with the $t=36$ h points being generally more isolated (i.e. distant from the matching $t=0$ neighbour) than the $t=24$ h points.

Initial PLS-DA modelling of the $t=0$ and $t=36$ h specimens using data integrated over 0.01 ppm. intervals and normalized to unit area showed reasonable class separation in component 2 (Figure 4c), but in the first component both classes were well intermixed, indicating that the major source of variation in the data was not attributable to storage condition. Together with the negative value of $Q^2$ for the first component, this implied that the metabolites related to group separation had a relatively small metabolite variance in comparison with the other common and overlapped signals, i.e. significant but very small differentiating peaks are superimposed on large common metabolite peaks. Further PLS-DA
analysis with full spectral resolution data, with or without orthogonal filtrations, confirmed the initial results using bucketed data, but gave a stronger model (small positive value for $Q^2_Y$) and elucidated multiple small storage time-related biochemical changes collectively underlying the inter-group differences.

The model loadings shown in pseudospectral format (Figure 5a) show highly significant baseline loadings regions at high and low chemical shift, indicating generalized but subtle alterations in protein structure which collectively contribute significantly to inter-group differentiation. Whilst the amplitude of e.g. the lipoprotein signal variations is large, it is not significant for class separation. The alteration in the protein profile can be attributed to changes from small peptide fragments, derived e.g. from multiple low-level protein degradation. The only significant spectral region that implicates a concentration change in a specific resolved metabolite was centred at $\delta = 3.26$ ppm. (Figure 5a) and assigned to trimethylamine-N-oxide (TMAO). The apparent decrease in TMAO may simply result from binding to the small peptide fragments liberated over the holding time, as noted earlier, effecting a reduction in free TMAO. A direct reductive mechanism for TMAO loss does not seem likely, as there is no observed alteration in the levels of trimethylamine, the primary reduction product.

Orthogonal filtration techniques were also applied to the data to assist the resolution of any groupings by removal of extraneous systematic variation unrelated to the primary metabolic profile.
to sample class, i.e. storage time. Two O-PLS-DA models for a 3 group discrimination illustrate that whilst the $t = 0$ vs $t = 24$ h group separation cannot be improved with the addition of extra orthogonal components to the models, the $t = 36$ h group can be progressively better differentiated from the other two groups by using further orthogonal filtrations. The discrimination for two 3-group PLS-DA models, using first $\Sigma = 4$ components, 2 aligned and 2 orthogonal (2,2 O-PLS-DA), and then $\Sigma = 6$ components, 2 aligned and 4 orthogonal (2,4 O-PLS-DA) are shown in Figure 5b and c, respectively in the 7-fold cross-validated scores, which have been plotted.

In addition to confirming the results of the integrated data, the full resolution O-PLS-DA model statistics provide additional information. For example the very small $R^2X_{\text{cum}}$ value observed indicates that the part of the total variance of the $X$ matrix (NMR data) related to the discrimination is very small, i.e. the difference in metabolite composition is very small, which is consistent with a split-specimen design with small differences between the two groups. These data confirm that a robust model separation of the $t = 36$ h specimen group from the $t = 0$ h group has been achieved. The reduction of the difference between $Q^2$ and $R^2$ with successive filtrations is exhibited in the 3-way O-PLS-DA model by the improved cross-validated scores separations of the $t = 0$ and $t = 36$ h specimens (Figure 5b and c).

**Discussion**

For the purposes of NMR-based analysis, storage of urine and serum for up to $t = 24$ h at 4°C does not detectably affect the metabolic profile. This result is consistent with expectations since the inter-individual variation was considerably larger than any effect of storage time. By $t = 36$ h, there were multiple small storage time-related biochemical changes collectively underlying inter-group differences between $t = 0$ and $t = 36$ h. Thus the UK Biobank protocol that stipulates shipment of specimens from the assessment centres at 4°C and storage at $-80$°C or $-196$°C within 24 h is robust for the analyses described, but time to freezing should not be extended to 36 h.

In this pilot Biobank study, it was noted that the specimen quality of the serum specimens was consistently high. In particular, there was no evidence of agglutination or precipitation in any of the specimens, even in the follow-up study of $t = 36$ h specimens. This contrasts with our historical analytical experiences with a wide repertoire of serum and plasma specimen types. For example, study participants with metabolic states ranging from latent through to overt disease such as severe dyslipidemic conditions, may produce plasma specimens which can have free lipids and/or extensively coagulated lipoprotein-derived fragments present. Amongst other effects, this can make homogeneity in aliquoting considerably more difficult to achieve, due to variable and non-representative sequestration of low- and high-molecular-weight metabolites into macro-assemblies lying in separated phases. Such 'lower quality' specimens are likely to be most strongly differentially affected by the protocol freezing process employed. It is thus recommended that a tightly observed and appropriate protocol be adopted for the freezing of all serum and plasma specimens, including rapidly freezing the sample to kinetically hinder the formation of ice crystals, which may puncture the lipoprotein outer bilayer. Although there were negligible systematic differences in specimen composition introduced by variable storage time, one of the challenging features of metabolic epidemiology, when ultra-large datasets are obtained, is that sub-classes of changes may emerge, with the result that corresponding sub-group analysis may reveal specific and idiosyncratic metabolic responses. These for example may originate from genetic, environmental, dietary, medicinal or even psychological influences such as stress. Detecting and modelling characteristics specific to such sub-groups is an ideal application of sophisticated multivariate methods to epidemiological datasets, and will result in improving the understanding of idiosyncratic metabolic sub-populations.

A singular advantage of NMR analysis of blood and urine is the minimal specimen preparation and disturbance prior to the quantitative spectral acquisition step. This is in contrast with techniques such as GC–MS, which requires pre-treatment of the specimen matrix, derivatization and then an on-column separation of the product (with the possibility for further on-column reactions) before any molecular characterization is made from the MS data. MS is matrix-dependent, and specimen-to-specimen metabolite quantitation may depend on the technology of the prior steps, including completeness of chromatographic separation. Questions relating to subsequent quantitation methods add further difficulties. Techniques such as GC–MS present a complex chain for assessing where metabolite-profile altering effects might occur, and thus any specimen variation observed may not accurately reflect the initial metabolic profile, and for example, derivatization may be influenced by speciation at the start of the procedure.

There are always molecular species that are difficult or impossible to directly detect by any given analytical technique. In GC–MS, the capability for suitable derivatization of non-volatile species is essential to the method. For $^1$H NMR, molecules with protons that exchange very rapidly with matrix protons (or which are without protons) will not be detectable. Furthermore, the intrinsic sensitivity of NMR, whilst continually improving, is not of the order of many other techniques such as MS or UV-VIS. Affinity assay methods are far superior for the detection of
e.g. hormones and low-level signalling species such as peptides, prostaglandins and enzymes. Many such species can be very precisely quantitated by the appropriate analytical method.

Intake of unusual dietary species, dietary supplements, pharmaceuticals and alcohol are often directly reflected in the NMR and/or mass spectra of biofluids, and may dominate the analysis if not controlled for, either by initial protocol compliance or by subsequent analytical exclusion. However, products of xenobiotics such as paracetamol carry important metabolic information in their own right and can, for example, hold clues as to activity or variation in enzymatic processes. Thus, in the present study involving nominally healthy volunteers, 25% of the cohort was revealed to have consumed varying dosages of paracetamol within the previous 24 h (Figure 3).

The PCA-based pairing procedure adopted here exhibited some capability for error detection, and we detected a labelling error among the samples. Human error is potentially an issue with epidemiological studies, especially where many personnel across dispersed laboratories collaborate to generate results from large population studies. In UK Biobank such errors are being kept to a minimum by computerized tracking of samples from the field to the biorepository store. We trialled but did not employ hierarchical clustering analysis as this may give differing results depending on the cluster algorithm scaling used.

Metabonomic analysis has the potential for an enriched derivation and interpretation of epidemiological data that may not be derivable via alternative methods, particularly if latent multivariate structures are involved. The scale of many epidemiological studies lends itself well to the capacity of multivariate analytical methods—which are fundamental to metabonomic technologies—to accommodate high-dimensional datasets. Modern multivariate methods include orthogonal filtration, correlation analysis and statistical correlation reconstruction techniques. Clinical chemistry and other measures such as time or age variables, class membership, etc, can be regressed against specific NMR spectral signals, and combinations of these may be cross-regressed in various ways. Data from a range of assays and analysis methods can be incorporated, with suitable scalings, into a ‘global’ fingerprint for model assessment, with models being predictively cross-validated to assure robustness. The multivariate approach is critically important to detection of effects where e.g. there is significance of weighted sums of assembled metabolites or assay results but not for any metabolite considered on its own. Such results may also highlight via correlations, important metabolic pathways derived from multivariate correlations within a set of spectra or by combining data across different analytical modalities. This scope provides a more ‘Systems Biology’ oriented approach to holistically uncovering the mechanisms that underlie metabolic differences.

Conclusions
In conclusion, the PCA-based analysis of reduced-resolution NMR datasets resulted in rapid identification of candidate split specimens. Group-systematic effects were indistinguishable between the specimens frozen after storage at \( t = 0 \) and \( t = 24 \) h at \( 4^\circ\text{C} \), using PLS-DA methods with or without implementation of orthogonal signal filtering, implying specimen stability. In contrast, by \( t = 36 \) h storage time, extremely low amplitude spectral differences over broad spectral regions characterizing peptide-like signals were observable between serum samples frozen at \( t = 0 \) and \( t = 36 \) h after collection. It is likely that these changes could be of larger magnitude in serum or plasma specimens which are of less than ideal quality, and that a rigorous protocol is needed for specimen archiving.

Working with the large specimen populations and the multiple metabolite measures available, metabonomic technologies offer an excellent future potential resource for enhancement of the richness and interpretable detail derivable from epidemiological studies.

Conflict of interest: None declared.

References


