Standardisation of Reporting Methods for Metabolic Analyses: A Draft Policy Document from the Standard Metabolic Reporting Structures (SMRS) Group

Editor and coordinator: John C. Lindon*

Section Coordinators: Glenn H. Cantor, Mark Earll, Nigel Hardy, Michael D. Reily, Donald G. Robertson, Susanna-Assunta Sansone

Contributors: Alexander Amberg, Henrik Antti, Dorrit Baunsgaard, Richard Beger, Stephen J. Bruce, Andrew Craig, Clare Daykin, John N. Haselden, Elaine Holmes, Hector C. Keun, Par Jonsson, Erik Johansson, Sunil Kochhar, Kerstin Kramer, Arno Krotzky, Lois Lehman-McKeeman, Johann Lindberg, Mahendra Navarange, Jeremy K. Nicholson, Jake T.M. Pearce, Robert Plumb, Jonathan Powell, Hartmut Schäfer, Ina Schuppe-Koistinen, Hans Senn, Manfred Spraul, Craig Thomas, Johan Trygg, Frans Van der Ouderaa, Elwin Verheij, Mark Viant, Ian D. Wilson, Svante Wold

*Please address any correspondence to this author at Biological Chemistry, Biomedical Sciences Division, Imperial College London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ UK. Tel +44 (0)20 7594 3194. Fax +44 (0)20 7594 3066. Email – j.lindon@imperial.ac.uk

(Affiliations and e-mail addresses for all contributors can be found in Appendix 1)

Version 2.3
09 March 2005

Executive Summary

- An initiative has been under way to investigate the reporting needs and to consider recommendations for standardising reporting arrangements for metabonomics/metabolomics studies.
- We describe here the formation of the Standard Metabolic Reporting Structures (SMRS) group and summarise the discussions that have been ongoing.
- We provide a draft policy document that covers all of those aspects of a metabolic study that are recommended for recording, from the origin of a biological sample, the analysis of material from that sample and chemometric and statistical approaches to retrieve information from the sample data.
- We also consider the various levels and consequent detail for reporting needs, including journal submissions, public databases and regulatory submissions.
- This document provides guidelines for such reporting and is intended to foster debate in the community.
- Some possible next steps are outlined.

1. Background to SMRS

Metabonomics is a rapidly expanding field with applications as diverse as pre-clinical drug safety assessment, disease diagnosis, plant metabolite profiling and environmental science, all employing various analytical techniques. With such a wide variety of data types, the need for standards to facilitate communication between different fields of activity, and to fulfil the needs of journal editors and regulatory agencies becomes clear. Among practitioners of metabonomics, a consensus has been emerging for a body of scientists to address the standardisation and reporting needs for metabolic studies. It is recognised that this effort is not a call for a standardisation of metabolic analysis study design, but a standardisation of information supplied by a metabonomic experimental protocol and subsequent report, such that the myriad factors that may affect metabonomic data are documented to allow proper scientific interpretation. As such, the results of this effort should be of widespread benefit to the community.

A first meeting of a group of interested parties, mainly based on personal contacts and current collaborations, was held at Imperial College, London, UK, in November 2003 to discuss the need for, and the mechanisms for, achieving standardisation of reporting of metabonomics and metabolomics studies. Similar definitions of metabonomics [1] and metabolomics [2] have been provided and here to avoid any confusion we use the term "metabolic analysis" throughout. There were representatives present from academia (Imperial College London and University of Wales, Aberystwyth), pharmaceutical companies (Pfizer, Lilly, Hofmann La-Roche, BMS and AstraZeneca), nutrition companies (Nestlé and Unilever), technology companies (Bruker, Waters and Umetrics), and from the US Food and Drug Administration (FDA). The group came to be known as the Standard Metabolic Reporting Structures (SMRS) group.

The second meeting was held in April 2004 at the Nestle Research Center, Switzerland, with wider representation, namely academia (Imperial College London, University of Wales, Aberystwyth, the Medical Research Council UK and the European Molecular Biology Laboratory – European Bioinformatics Institute), pharmaceutical companies and related contract organisations (Pfizer, Lilly, Hofmann La-Roche, Bristol-Meyers-Squibb, AstraZeneca, GlaxoSmithKline, Aventis, NovoNordisk and TNO Pharma), nutrition companies (Nestlé and Unilever), a plant science company (Metanomics), and technology companies (Bruker and Umetrics). Other organisations who were invited but were unable to attend, were Novartis, Waters, Procter & Gamble, Merck, Sanofi-Synthelabo, Lipomics Inc., and the US National Center for Toxicological Research (FDA).

The minutes of the two meetings and the agreed action points are summarised in Appendix 2 of this document.

The main proposal was to prepare a draft policy document and possibly to submit an abridged form to a journal such as Nature Biotechnology by the time of a meeting in late 2004. It was also proposed that a letter be sent to journals interested in standards and reporting structures for their feedback. The success of the MIAME protocol for gene array data [3] was recognised..

The full discussion document and other relevant documents will be posted to a new web page devoted to the SMRS group, registered and hosted from Imperial College, (www.smrsgroup.org) and an e-mail circulation list has also been developed to keep all interested parties informed

2. Interactions with other groups

(Originally contributed by Susanna-Assunta Sansone)

The SMRS group is aware of others interested in standardisation of data sets and methods of reporting in "omics" technologies. The SMRS group welcomes interactions with other efforts that have begun to evolve for the exchange of standards. To capitalise on these efforts, representatives of the SMRS group are also directly participating in some of those initiatives, fostering interactions and laying the ground for further collaborations. Independent initiatives are regarded as important, because they target specific requirements for the particular omics technologies being used and they understand the particular context of the fields of application (e.g. pharmaceutical, medical or environmental). Some regulatory bodies, such as the FDA and Environmental Protection Agency (EPA), have also published their policy or guidance on genomics data submissions [www.fda.gov/cder/guidance/5900dft.doc, www.epa.gov/osa/genomics.htm]. There will be a fundamental difference in both the design and objectives of the efforts focused on regulatory submission of data versus those focused on needs of the research community. The former aims to accelerate the review process, facilitate proprietary data submission and optimise data visualisation. The latter aims to develop detailed, highly structured databases, to facilitate data exchange and to provide a fundamental biochemical understanding. However, there may be a value in developing a compatibility between regulatory and research objectives in the design of these data standards. It is a generally accepted view that duplication and incompatibility should be avoided where possible, maximising the synergy and optimising harmonization. In particular, a unified approach to describe and report the biological component of an experiment that is common to different "omics" technologies (trascriptomics, proteomics and metabonomics/metabolomics) or disciplines (e.g. pharmacogenomics, toxicogenomics, environmental genomics) is strongly recommended. Undoubtedly, specialised information is needed by certain types of application, but a unified model should be able to encompass them all.

A summary of standardisation efforts within the "omics" and related communities with particular focus on toxicological applications is provided in Appendix 3. This includes initiatives that specifically address data format and reporting structures for the exchange of information and database applications and discussion forums aiming to a broader understanding and acceptance of "omics" data.

3. Some General Considerations for any Data Set Compilation

(Originally contributed by Nigel Hardy)

3.1. Data types

The compilation of metabolic data sets comprises two aspects, each analytical data acquisition produces numerical values as estimates of metabolite concentration and these estimates vary in data structure between different analytical procedures; all such estimates are carried out on biological material (which has a provenance) and in a particular study context and this produces a wide variety of supplementary data (meta-data, or context data). Both of these aspects are crucial to any interpretation and therefore both are required as part of data set compilation.

The structure of the result of a single instrument analysis to estimate metabolite composition can vary significantly, depending on the analytical technique and subsequent data processing. The (unattainable) goal is to have a complete list of the absolute abundance of all fully identified compounds. This can be termed a full metabolite *peak table*. Some analytical techniques produce such peak tables, but never full tables for typical biological samples. Other techniques do not produce peak tables at all. Rather they produce some reasonably reproducible pattern, which reflects the metabolite composition. This can be called a *fingerprint*. A characterisation of this range of results has been provided [2] and results for each analysis must be accompanied by clear meta-data (annotation) to describe the nature of those results.

A peak table has two columns: a list of identified compounds with associated abundances. The compounds will be identified with varying precision and terminologies. Assuming that precise identification is possible, the naming convention must be identified. IUPAC [www.chem.qmw.ac.uk/iupac] and KEGG [www.genome.ad.jp/kegg] are examples of possible "authority" for names but others exist and each has advantages and disadvantages. Names for compounds must all be related to some authority to give them meaning. Precise identification will not always be possible. A compound may only be identified to its chemical class (oligosaccharide, for example), but still be a recognisably distinct peak, found reliably. A locally generated descriptive name may therefore be used. Still less precise is a local name, such as "Peak42" which reflects no understanding of its chemistry but can be confidently distinguished. All such local names mean again that the authority must be stated – in these cases, it may be a laboratory or organisation.

The abundances will be relative or absolute. Absolute values require units and relative abundances will require a context, such as relative to the most abundant compound, to the total amount of all compounds detected, or to an introduced standard compound for example. All compounds in one peak table will have abundances of the same type.

Fingerprints are typically 2 or 3 dimensional data. For example, FT-IR analysis generates a typical 2-D fingerprint (intensity against wave number). On the other hand, GC-MS generates a 3-D fingerprint (intensity against m/z against time) and units will be required for each axis. The fingerprint will typically have been generated by numerical routines after the initial collection. These routines and their parameters, such as binning and resolution issues, must be reported, for comparability of the data to be established.

In multivariate data analysis, peak intensities or one of the discrete axes from 2-D data fingerprints are the variables or attributes. They are the "X variables" in data mining contexts.

The dates, instruments and operators involved in analytical data collection should be recorded and are necessary to account for systematic variance potentially introduced by the process. Post instrument/pre-analysis processing details may also introduce bias or otherwise affect the nature of the estimate.

3.2. Sample contexts and classes

The source of the biological material must be identified. Full details of material origin and species, and of its maintenance must be available. These are typically species specific and the authority for all identifiers will be necessary to ensure comparability. For examples from plant-derived materials, details of the growth conditions are required. Again, those aspects of the environment shown to cause greatest variance (noise) will vary between species. The natures of the sample (whole organism, organ, tissue etc.), together with the experimental protocol for its collection are required. Vocabularies or ontologies appropriate for the species must be cited to qualify the descriptions.

All sample storage, handling and preparation must be recorded at a level that allows comparability to be established. This will include a range of activities (particularly for solvent extracted samples and in the case of GC-MS), which will bias the metabolite complement estimate in characterisable ways.

The experimental design must be reflected in the context. Treatment v. control; time series; treatments of variance (e.g. Latin squares) must be accessible to permit correct compilation of data sets.

All these data provide the opportunity for selection of appropriate data sets and/or to provide "Y attributes" or classifiers for data analysis.

3.3. Sample and measurement technique selection

Each analytical data collection can contribute a set of variables, but a given data analysis may not be based on all of them. A justified rationale for selection must be provided and the details of the selection recorded. Additionally, a particular analysis might be based on variables from more than one type of analytical technique. Where these can be shown to be for the same sample, this is entirely appropriate. Through concatenation of such sample analyses and selection of variables, a set of variables for the analysis can be assembled and that process can be reported. This can be characterised as joins and restrictions in relational terms and defines the columns of a multivariate data table.

Which set of spectroscopic estimates to be included as rows of the table must be defined by reference to context data and can be characterised, in database terminology, as relational restrictions or selection of matches. Typically, strictly comparable estimates from a single trial or experiment will be included.

Additional columns of the data table will be derived from the context data. Most obviously, class variables for supervised methods will be included. In other contexts, the values for factors in traditional experiment designs will be included to permit classification of the estimates (e.g. for MANOVA or similar procedures).

3.4. Vocabularies

Much of the data described above requires controlled vocabularies and establishment of these can be difficult. There are often pre-existing competing vocabularies in a field, broadly comparable vocabularies in different fields and problems with changing and developing lists. In general terms, it seems expedient to only require that a value and its *authority* (i.e. the vocabulary from which it is taken) be provided. This has several advantages. It prevents accidental incorrect comparison, both through a value

being taken to have the same meaning in two cases where it does not and the assumption that two different values have different meanings. It permits development of vocabularies, though the hierarchical definition of vocabularies including both public and local lists. It can permit semi-automatic establishment of comparability. It should tend to foster agreement on vocabularies commonly used in reporting results. The discussion above is based on this underlying model. That does not preclude the creation and identification of particular vocabularies as "best practice" in specific situations. As an example, in toxicological studies on rodents, a consistent terminology could be defined for describing the histopathology findings across different laboratories. However it is recognised that the more wide-reaching a nomenclature becomes, the more difficult it is to implement. For example the SEND Consortium which tried to harmonise all meta-data did not develop a uniform glossary of terms for pathology.

It may be noted that defining reporting structures in terms of external vocabularies is not avoiding the issue. Rather it is separating it and allowing independent work. Discussion of data structures and of vocabularies can proceed independently. The structure is not invalidated if a vocabulary changes or is superseded. Necessary vocabularies have been implied in the discussion above.

4. Biology reporting items

(Coordinated by, and originally contributed by Glenn Cantor and Don Robertson)

4.1 Range of use of metabolic analyses

Metabolic analyses have been used in animal, plant, and microbial systems to describe and understand a variety of biological effects, including response to environmental factors, diet, disease states, toxins, or pharmaceutical agents, enhanced understanding of the results of increased or decreased gene expression, and taxonomic studies. Although the approach has been used for the past 15 years, the advent of more powerful chemical analytical equipment and techniques and enhanced chemometric methods have led to much more widespread use in recent years.

Changes in the metabolite profile detected in biofluids such as urine and plasma, or at a tissue level, are the result of very complex regulatory mechanisms including changes in gene and protein expression levels, as well as physiological changes in response to external stimuli, disease, ageing, nutritional and other environmental changes. By profiling changes in metabolite composition, either in tissues or biofluids, following intervention by pharmacologically active compounds, it is possible to generate information about the overall systems response to such challenges. The integration of biochemical information with proteomic and genomic data, as well as traditional pharmacological and toxicological endpoints, might provide a better understanding of the response to external stimuli and ultimately may facilitate the selection of better drug candidates with a higher probability of success and in a shorter period of time. The main areas of application of metabolite profiling throughout the drug discovery and pre-clinical development process are target evaluation, lead generation, lead optimization and safety evaluation of drug candidates.

This has led to a proliferation of metabolic analysis techniques, a healthy development which has fostered innovation and creative development in the field. The downside of the broad acceptance of the technology is that various protocols,

methods and reporting structures have developed within sub-disciplines employing the technology. It is also important to coordinate these standards and methods of data reporting among groups addressing similar issues in genomics and proteomics and with efforts to improve electronic data reporting. This section deals with generating and interpreting the biological data that is part of any metabolic experiment.

4.2. Role of metabonomics in pharmaceutical R&D and in the clinic

Metabolic analysis can be used to characterise animal models of disease and efficacy as well as to assess the consequences of target modulation in vivo and in vitro to evaluate specificity, efficacy and potential on-target toxicity of novel candidate therapeutics. Identification of biomarkers of efficacy early in discovery is also a highly desired objective. Screens can be performed *in vivo* using specific strains or transgenic and knock-out animal models of disease in which, metabolic analysis offers a way to assess the global biochemical effect of a new chemical entity (NCE). Even more advantageous is the ability of metabonomics in such models to highlight potential novel targets and provide a platform for evaluation of the target. Other organisms or cell systems being assessed within the pharmaceutical industry include embryonic stem cells, xenopus, zebra fish, *C. elegans*, *Drosophila*, *S. cerevisiae*.

Lead generation includes characterisation of chemical series that have been shown to affect the target of interest. Limited *in vivo* or *in vitro* experiments can be used to demonstrate efficacy and flag potential toxicity issues in order to facilitate selection of suitable structures.

One of the clear advantages of "omic" technologies in pharmaceutical discovery lies in the potential to address both safety and efficacy endpoints, potentially within the same study, thus facilitating the selection of drug candidates by decreasing the probability of early attrition. These efforts may include predictive toxicology approaches using chemometric models based on a training set of well-characterised compounds.

The main aspects of applying metabolic analysis at this stage are investigations of mechanisms of toxicity, cross-species extrapolation and identification of non-invasive markers of toxicity that have the potential to be evaluated in the clinic. It is important to note that it is unlikely that metabolic analysis in isolation will provide the answers needed to investigate mechanisms of toxicity. Integration of metabolic analysis data with traditional, well characterised toxicological endpoints is essential. Furthermore, a systems biology approach integrating genomics, proteomics and metabolic analysis data with toxicological endpoints may not only have a better chance of elucidating a toxicity and provide a perspective concerning its relevance to man, it may also provide a mechanistic association for biomarkers with the potential of being applied in the clinic.

It is perceived that an area of great opportunity for metabonomics to have an impact is in the clinical arena. For the purposes of metabolic analysis standards, clinical studies can be broadly described as falling into three categories defined by the size of the study required. These are large-scale 'epidemiological' studies which may look at disease sub-populations in search of disease-specific profiles/fingerprints that enable either an earlier diagnosis or biochemical changes which point towards new target/mechanism of action. Small, early Phase I clinical 'drug development' studies

in volunteers (not patients) whereby new chemical entities (NCEs) may be assessed for safety and tolerability and moderate size to large trials in which NCE's are tested in patient populations with the inherent difficulties that this implies.

Implicit in all clinical studies, regardless of size, is the definition of the metabolic profile of a 'normal' population (see above) so as to be able to infer relevance of any observed metabonomic changes. This is very difficult in large scale 'epidemiological'-type studies which need to be powered accordingly to negate impact of the variation in humans. However, such variation can be better controlled in residential volunteer trials whereby the population is tighter in terms of metabonomic or biochemical variability (usually dominated by diet and exercise) which make these ideal studies for defining baseline variation or 'normality'. Assessing metabolic changes in patients entails a little of both, in that not only must the metabolic differences induced by routine variables such as diet, and exercise be understood (and controlled if possible), but metabolic variation induced by the disease process under investigation is as important. For example it might be easy to spot an inflammation metabolic signature in a control population, but much tougher in a population of COPD or asthma patients.

4.3. Submissions to regulatory organizations

To date, very little multi-variate metabonomic data has been submitted to regulatory agencies although metabolic biomarkers have been used extensively. The perception is that the FDA has been exposed to some metabonomic data (related to preclinical vascular toxicity) but the validity of this new type of data will require extensive investigation at NCTR specifically and in the FDA in general. This highlights the very real need for non-agency experts to work with and support the agency scientists so that they can fully appreciate the utility of metabonomics. In Europe there are ongoing initiatives through EfPIA to educate preclinical assessors on "omics" data in general. Currently there is no steer on what the Japanese regulators will expect or what interactions they have had to date (none known or reported thus far) but they will probably take their lead from the FDA.

4.4. Common factors across metabolic analysis studies that need standardisation

In order for a metabolic dataset to be put in perspective, the relevance of any changes observed should be put in context in reference to the test system normal or control population. As with any "omics" platform there is inherent variation within the data (defining the boundaries of 'normal') and any such variation will be determined by many factors such as species, strain, age, diet and test facility.

To facilitate a standardised approach to the reporting of metabolic data, there are criteria or test system information that should be apparent in any textual/tabular data presented. Below are highlighted a number of factors that determine the boundaries of metabolic 'normality'. An attempt is made to put into context the impact of these factors as they affect different broad areas of metabolic investigation.

4.4.1. Phenotype and genetic modification

Non-clinical: Clear definitions of animal species, strain, gender, age range, weight range, wild-type/ background strain, strain designation (using universally accepted rules) as applied to knock-outs, transgenics and/or disease models.

Clinical: Definition of ethnic background, age range, weight range (BMI) or other standard inclusion criteria (including disease information). Pharmacogenetic data may exist and if allowable should be included (often if anonymised).
Plant/others: Clear definition of species, strain, age range, wild-type/background strain, strain designation and stability (using universally accepted rules) as applied to knock-outs and transgenics.

4.4.2. Dietary control
Non-clinical: Statement of frequency of food (*ad* libitum or restricted), time of feeding, diet type, batch ID, manufacturer, and where possible carbohydrate, lipid and protein relative concentrations. For reporting of preclinical data, every effort should be made to ensure common batches of diet used throughout. Batch variation can happen through the seasons and often relates to differing quantities of fishmeal protein blended into standard animal chow.
Clinical: Definition of time of last food intake relative to sampling, exclusion criteria (certain foodstuffs) and state any specific dietary anomalies. Unless clinical studies are specifically designed, then beyond the standard exclusions it is often very difficult to control human dietary intake beyond the standard exclusions. This can have a huge impact on the profile, so ideally any bespoke metabolic investigations should attempt to standardise diet as much as is feasible to minimise variation.
Plant/others: Nothing beyond those items considered under environmental factors.

4.4.3. Environment
Non-clinical: Definition of the test facility conditions light cycle period (*e.g.* 0700-1900hrs), humidity (35-75%), water source (distilled, tap, bottled, etc., with purity assessment), sample collection times with respect to the light cycle.
Clinical: Controlled early trials may have standardised conditions that should be defined, otherwise this will be very difficult to control and define.
Plants/others: Definition of the test facility conditions light cycle period (*e.g.* 0700-1900hrs), humidity (35-75%), water source, gaseous environment.

4.4.4. Housing
Non-clinical: Definition of caging conditions (single *vs* group-housed), types of cage used (size) and methods of sample collection.
Clinical: None beyond those considered under environmental factors.
Plants/others: None beyond those considered under environmental factors.

4.5. Summary
It should be clear from the previous discussion that the state of biological standardisation for metabonomics experiments is currently non-existent (although these may well be in place within them). Most of the required standardisation is common sense and it is likely that adopting a "uniform" approach should not be impractical, as it is defined by "best practice". The past few years have shown that no matter how robust and precise an analytical platform, when dealing with systems biology, the meta-data are as important as the omic data in understanding and interpreting results. In the past the solution to ignorance of systems responses to environmental, physiologic and xenobiotic stimuli was to increase the "n" size within the experimental groups hoping for an "average" response suitable for making some type of interpretation. It would however, be advantageous if an understanding of idiosyncratic reactions (of low frequency of occurrence) could be achieved.

Metabolic analysis and system biology approaches may generate that understanding. Much of what has been ascribed to normal variation is probably attributable to some definable cause and hence controllable. The ability to understand these factors will not only simplify data interpretation, but may lead to decreased animal utilization as smaller group sizes become feasible. It should be made clear that this is not a call for a standardisation of metabolic analysis study design, but a standardisation of information supplied by the metabonomic protocol and report such that the myriad factors that may affect metabonomic data are documented to allow proper scientific interpretation. As such, this will be of widespread benefit to the community.

5. Analytical Technique Reporting for Metabonomics

(Coordinated by, and initially contributed by Michael Reily)

5.1. Sample handling

The purpose of this section is to provide guidance on analytical aspects of reporting metabolic data. Since this is a rapidly expanding field, in which new technologies are rapidly being applied, one cannot anticipate all issues relevant to every potential technology. Rather, the following should be general enough to allow for incorporation of new approaches as they come into play.

Another important issue is the importance of proper sample preparation. Sample preparation will be dependent the particular technology being used and the questions being asked. For the purposes of this document, sample preparation is considered to consist of an audit trail of what has been done to the sample from the time of collection to the start of data acquisition. At a minimum, factors such as, buffers, storage conditions, extraction protocols, reagents used, internal or external standards, and derivatisation methodologies should be described in detail. If robotics techniques are used to prepare samples, details of the methods and types of equipment used should be provided.

Descriptions of sample handling during the data acquisition should cover details such as the sample container (type and source), storage conditions (temperature, length of time), method of delivery to the analytical instrumentation (e.g. flow injection, HPLC, individual sample containers) along with the types of sample changing equipment and software used and the temperature and the time period the samples are stored on the instrument. If flow injection methods are used, carry-over assessment and sample dilution should be reported. Recommended procedures for NMR spectroscopy and sample handling for high throughput, general metabolic analysis of urine or serum samples, are given as examples, rather than recommendations, in Appendix 4. This also provides some procedures for preparing extracts of tissues for NMR spectroscopic studies.

One aim with using GC- or HPLC-MS in metabonomics is to analyse as many compounds/metabolites as possible with as high accuracy and precision as possible. To be able to do that, the sample preparation steps must be optimised; maximising metabolic information at the same time as biological activity and analytical errors are minimised. The wide range of metabolites occurring in biological samples makes it impossible to find the “best” settings for all of the metabolites present, so compromises have to be made. Design of experiments (DOE) and multivariate

evaluation can be key methods in this area to construct robust sample preparation protocols for GC- and LC-MS analysis in metabonomics [4].

5.2. Data acquisition and instrument level data processing

Firstly, reporting of some form of validation of the analytical equipment and the specific processes for the data acquisition should be included. It is assumed that proper analytical method validation [5] was used on the instruments and methods generating the raw data. All analytical methods used have to follow general guidelines of Equipment Qualification (EQ) and Design Qualification (DQ), as established by ISO9001 procedures. Other standards, such as Installation Qualification (IQ) achieved with specification tests, Operation Qualification (OQ) and Performance Qualification (PQ), achieved by such procedures as GLP (Good Laboratory Practice) and / or GCP (Good Clinical Practice) might apply in appropriate circumstances [6].

It is important to define what is meant by "raw data" for the analytical techniques used, mainly NMR spectroscopy and MS. In NMR, raw data, typically the free induction decay, comprises a set of digitally sampled data points. However, this has no real interpretable meaning without extra data processing, which therefore should be specified in a fashion similar to that for recommendations for the presentation of NMR structures of proteins and nucleic acids [7].

Next, the analytical technique hardware used should to be specified. For NMR spectroscopy, such parameters include the manufacturer, model, software, magnetic field specified by the resonance frequency of a particular nucleus, usually ^{1}H, and the type of electronics console and a description of the probe head has also to be specified. The size and nature of the sample tube or container has to be described. In case of novel experimental protocols or home-built systems, additional details have to be provided for adequate description. For all analytical approaches, consideration should be paid to randomisation of sample running order, use of replicate samples, and any instrumental parameters relevant to the particular approach. Thus for NMR, the shimming method, sample pH, sample temperature, sample volume and probe sensitivity should be adequately described. For MS, ionisation mode (+ve or –ve), tuning, source, cone voltage, gas flows and probe temperature would be required.

Secondly, the protocols for raw data acquisition have to be given. Minimum requirements of any analytical platform dependent processing to be reported include the software used, including version and supplier (including the version of the instrument software). For NMR spectroscopy, either a diagram of the pulse sequence and / or an ASCII file with the pulse program code has to be shown or an adequate reference to previously published routines has to be given [8]. Details of the raw data (size of the time domain, dwell time), timing parameters (pulse lengths, delays, e.g. recycling delay and total time for data acquisition), r.f. power levels or equivalent 90° pulse duration should be provided. If necessary, phase cycling, pulse shapes and magnetic field gradients used should be described. Raw data and protocols for their acquisition should be stored together [9]. Appendix 5 shows some examples typical of NMR spectroscopy for biofluid studies.

Quality assessment of analytical data should be described, including stability of analytical procedure, e.g. drift, use of standards (this applies to both NMR and MS), error estimation of data reproducibility, reporting of excluded data due to

spectrometer errors and their reason for exclusion. A description of analytical instrument calibration should be provided, with appropriate dependence on the technology utilised. For NMR spectroscopy, a description of the range of internal standard line shapes or line widths observed across all spectra in a study should be reported along with an indication of the quality of water suppression. For MS, resolution, sensitivity, retention time alignment, mass calibration and mass accuracy should be reported.

In addition to the above, when GC- or HPLC-hyphenated methods are used to introduce samples into the spectrometer, chromatographic parameters such as mobile phase composition, column type and dimensions, particle size, gradient profile, flow rate should be included.

Also, instrument level processing could be performed using a different software package to that supplied by the instrument manufacturer, and if so, this should be reported. Details of any analytical data processing such as for e.g., NMR spectroscopy, zero-filling, linear prediction, convolution with a window function, Fourier transformation, phase and baseline correction and calibration, and for MS and any other analytical techniques, each processing step applied to the raw data must be reported. For MS, descriptions of applicable parameters and procedures for centroid or profile data, background subtraction, signal averaging, deconvolution, spike reduction, removal of adduct ions, charge deconvolution, etc. should be provided. Nomenclature should follow authorised guidelines, as published for NMR spectroscopy [10,11]

5.3. Post-instrument data pre-processing

This section describes the steps taken to prepare the data taken from analytical instruments for univariate or multivariate analysis. The objective is to ensure that adequate descriptions of such pre-processing information are provided to ensure that results can be reproduced in independent studies or compared across studies.

For NMR spectroscopy, descriptions of applicable parameters and procedures such as for further baseline correction or integral limits for data binning should be provided. Similar parameters would be needed for MS studies.

Often in chemometric analysis of metabolic spectroscopic data, a binning procedure with inevitable loss of precision has been used. However binning, i.e. dividing of spectra or chromatograms into regions or windows, does not lower the resolution of the raw spectral or chromatographic data. The binning is merely a way of facilitating multivariate analysis for comparison of large sample populations characterised by a vast number of highly correlated variables. Once interesting regions or features are identified, the original high resolution ('raw') spectra may be inspected for detailed analysis.

The required data formats will depend on the type of analysis being reported, e.g. quantitative metabolite identification or multivariate data analysis. For quantitative metabolite identification, molecular identification, peak intensity, concentration standard, concentration of analyte (absolute, relative or corrected for dilution), normalisation method and error analysis should be described. Unknowns should be uniquely identified with a unique descriptor (e.g. retention time and mass for HPLC-

MS). For multivariate analysis, the procedure for data reduction (e.g. binning, peak picking) should be described along with excluded spectral regions, alignment approach for peak picked data (unaligned data should also be reported in this case).

It is recommended that for quantitative metabolite identification a table with the information described above should be provided. Unknowns should be identified with some spectral indicator. For example an unknown measured in an NMR spectrum could be identified as Uxy(Y), where U indicates an unknown, x is the chemical shift of the measured peak, y is the multiplicity of the peak (s for singlet, d for doublet, t for triplet, m for higher multiplets), and Y is the sample type (plant extract, tissue, biofluid, etc.). Using this format, an unknown urinary component with a doublet NMR peak appearing at a proton chemical shift of 2.34 ppm would be designated as U2.34d(urine).

A critical issue with mass spectra is their alignment and resolution. For multivariate techniques to work, a data table must be constructed so that each observation is characterised by the same number of variables and that each of these variables is represented across all observations. Mass spectra of metabolites typically have very few peaks and a large number of essentially zero variables and are thus relatively sparse compared with other forms of spectroscopy. This can complicate multivariate modelling, since fairly similar chemical structures may be projected into very different areas of multivariate score space. A number of possible transformations of the mass spectra exist that may be used to extract relevant information from the data [12]. Higher resolution mass spectra may be presented by recording the most prominent exact masses found at the apex of each peak in the chromatogram. In this way the 3D HPLC-MS data is unfolded so that variables for analysis have the paired form “retention time;mass”. In this case great care needs to be exercised in the operation of the chromatographic system in order that different samples (observations) do not drift in retention time.

The method for thresholding peaks in mass spectrometry should be reported. In addition, the method used for aligning equal mass values is of critical importance and should be described, for example based on an adjacent mass unit. Careful calibration alignment and data processing is needed to ensure errors are not included in the data and this should be noted.

Mass spectra are often presented as normalised to the biggest peak, usually the parent ion. From a data analysis point of view normalisation radically alters the correlation structure of the data [13]. The problem is understood by realising that if one variable increases all the other variables will decrease because the sum is fixed. The consequence of this is that spurious negative correlations may be found between large variables and positive correlations among the smaller variables. The raw un-normalised data may be an alternative source for chemometric analysis. Using an internal standard is another way of overcoming the problem of normalisation in some cases, but this may not be desirable for some samples. Details of any such manipulations should be logged.

Chromatographic data have further variability in that different chromatographic runs may be affected by drift in the elution times of peaks. Alternatively chromatographic data processing and integration capabilities can be used to produce a peak table. In

this way drift free data is obtained for analysis. Additional information provided by the system, for example Diode Array UV data, may be used to check the peak assignment. This information would not be apparent from analysis of the raw amplitude versus time chromatogram data outside the instrument.

The chromatographic alignment problem may be more severe for HPLC-MS than GC-MS, which is considered a more reproducible method. However, within metabonomics very similar samples are often compared, therefore most, if not all of the samples will contain many of the larger peaks observed. The total ion chromatograms can hence be aligned if need be, e.g. by finding the maximal covariance between the chromatograms [14]. The largest peaks in the chromatogram must be found in all measurements, because they contribute most to the covariance. Due to the similarity of samples within most metabonomic studies, this method maybe a desirable one.

One method of handling 3-dimensional GC- or HPLC-MS structures is to summarise the data in the chromatographic direction by a combined mass spectrum for each individual sample. Again, the methods used should be recorded. An alternative way of analysing the data is to divide the chromatographic dimension into specified retention time windows and then, on an individual sample basis, summarise the profiles in each window by its mass spectra.

Probably the most powerful but also the most difficult method for dealing with 3-dimensional GC- or HPLC-MS data structures is to deconvolve the data by applying some type of curve resolution technique. Hence, in the ideal case, all components in the data would be resolved so that all metabolites would be represented by one unique chromatographic profile with one corresponding unique mass spectrum. This procedure can be performed for all the individual samples separately and the spectral profiles from the resolved components are then used for matching of components between the individual samples.

6. Chemometrics reporting for Metabonomics

(Coordinated by and originally contributed by Mark Earll)

6.1. Post Instrument Processing

The accepted convention with multivariate analysis of metabolic data is to compile a table in which each sample or observation is represented by a row of the table and each variable or descriptor defines a column of the table. By this approach the precise meaning of each row and column of the data table can be unambiguously understood. It should be clear how the sample set derives from the experimental design used, by appropriate reference to the biological context or meta-data already discussed. The use of clear labelling schemes with separate columns for each observation attribute (meta-data) such as "Treatment", "Dose", "Time", "Animal No." at the start of the table is recommended.

Every process used in the transformation of the raw analytical data to the final data table prior to modelling should be fully recorded. This should also include a description of any automatic processing performed on the raw instrumental data by the software controlling the analytical device. As a minimum these processes should be stated along with the order in which they were applied.

The next stage is manipulation or re-arrangement of the processed data which may then form the table used for data modelling. In the case of NMR, some regions of the spectrum may also be excluded from further analysis and the remainder of the spectrum may be further processed by binning or some other procedure. For LC-MS approaches this might include estimation of concentration of species from the analytical data. Where appropriate for each quantity derived from analytical data, such as metabolite concentrations derived from LC MS data, an estimate of the uncertainty in the quantity should be made.

Finally, there are a group of transformations that may be applied immediately prior to the data modelling stage which are largely independent of the analytical platform used which need to be recorded. In all cases the object of this stage is to enhance information recovery from the analysis of the data table. It is possible to classify these transformations as (1) row operations which include processes such as normalisation or standardisation *i.e.* rescaling of each observation (row) with respect to some quantity so as to make individual observations more directly comparable, (2) column operations that are performed on each column individually, such as centring and scaling with respect to range or variance, and (3) global operations using parameters derived using the entire data table simultaneously. All of these operations form part of the recommended audit trail.

6.2. Methods of Analysis

Having defined the initial processed data table, it is important to state the objective and type of the modelling to be conducted. The major distinction is between supervised objectives, such as regression or classification, and unsupervised, such as clustering or visualisation.

Supervised methods use some prior knowledge of the data, which can be considered as a new data table 'Y', to derive a model of the original data table "X" from which Y can be predicted (i.e. class membership). The aim of unsupervised methods is to derive a model that captures some intrinsic structure within X, (i.e. trends, groupings and clusters) without any previous knowledge of those relationships.

In reporting the modelling process, it is necessary to state which algorithm was used to derive the model, any specific parameters relevant to the algorithm and the particular implementation of algorithm used. In the case of proprietary software where the details of the implementation cannot be revealed, the software version and vendor must be provided. The level to which the model itself should be reported is very case-dependent, and often in the case of supervised methods the focus is on the predictive validity of the model. For purely exploratory work the standards of integrity are the same as those expected for any visualisation in scientific reporting: proper labelling describing scale and other manipulations should be given.

The level to which outliers or other exclusions must be reported depends on the reasons for the removal of any observations. However, in principle, every sample not utilised from the original experimental design must be accounted for and justification given. Of particular concern are those outliers detected during model construction that are removed to 'improve' the model. It is preferable to have external justification for such exclusions, although it must be recognised that such actions are often part of the normal, iterative process of data modelling. A well-implemented auditing system

allows each model iteration to be scrutinized retrospectively, and facilitates data-mining for the root cause of anomalies. As a minimum for reporting, the model in which an outlier was detected should at least be described, and reasons for the deletion given.

6.3. Recording Validation of Chemometric Methods

Many different chemometric and data modelling methods may be applied to metabolic data but broadly they can be categorised in terms of the objective of the analysis - overview and data visualisation, classification and regression (or calibration). Validation methods will also vary with the objective and the precise algorithm or method used, but certain minimum standards of validation must be adhered to.

When modelling data, it is often possible obtain a model which fits the data well showing apparent trends and patterns. If significant noise is found within the data the model may become over fitted (modelling noise) and if the experiment is repeated, the predictive ability will suffer, as the new set will have a different noise component. It is therefore essential to use and record methods that can distinguish the regularities in the data from noise in order to prevent over-fitting and to give a realistic estimate of the predictive capability. Methods such as internal cross-validation where a portion of the training data is left out of the model and then predicted are helpful in this regard.

The real test of predictivity of a multivariate model is to predict what will happen for future observations. To do this, it is necessary is to obtain a dataset not available at the time the model was built (prediction set) and apply the model to this new data. In assessing the model in this way it must also be realised that instrumental drift and experimental parameters may change so it is important to describe the experimental procedure, equipment and personnel as fully as possible. A glossary of definitions of some of the terminology used in chemometrics is given in Appendix 6. The success of the data overview, classification or regression may then be assessed.

Sometimes it is found that the model predictivity is inadequate, because of new signatures in the prediction set not present in the training set. This implies that the training set must be updated with this missing information and a merger of the training set and the prediction set is therefore likely. As a consequence, a new, updated model must be developed and a new prediction set must be selected in order to pave the way for predictivity assessment of the updated model. This means that model derivation is usually an iterative process and many repetitive cycles of model predictivity assessment is needed. All of this should be recorded in the audit trail.

In cases where new data are not available, the dataset may be divided into test and training sets and the model that is built on the training set, is used to predict the test set. This may be repeated, swapping the test and training sets and the average of the estimated predictive ability between the two (or more) sets taken as a measure of model success. In cases where data are split into test and training sets rational ways of selecting the training set in order to avoid operator bias should be chosen.

6.4. Model Overview and Data Visualisation.

As the objective in data visualisation is of a descriptive nature, validation standards are not as rigorous as for class or value prediction studies for example. The

following should be reported - the amount of variance explained by the data display, the type of normalisation, filtering, centreing, scaling, and transformation of data and display parameters. For a valid model, new data (prediction set) or held-back data (test set) mapped into the model space should show essentially similar trends groupings and patterns.

6.5. Classification and Regression Methods

Classification methods may be divided into two types, *unsupervised* and *supervised*. In unsupervised methods no prior information about the classes is included within the data to be modelled, and the data are not pre-biased. Techniques such as principal components analysis (PCA) fall into this category and may be used for the detection of groups or classes within data. In supervised classification, the investigator chooses the training set based upon prior knowledge of biological or metabolic significance.

As supervised classification is an inherently biased technique, validation by means of an external test set is mandatory to judge the validity of the method. This can be using new data (prediction set) or held-back data (test set) mapped into the model space which should show essentially similar trends groupings and patterns, and a measure of classification success when the model is applied to an external test set should be supplied. Also, a measure of model confidence appropriate to the data distribution should be given. Measures of classification success may be by means of confusion matrices (detailing true positive, true negative, false positive, false negatives, or Cooper statistics) or percentage classification success rates.

Regression involves the prediction of one or several continuous Y variables. The model is built upon a set of calibration standards with accurately known Y value (often a chemical concentration). It is recommended that new data (prediction set) or held-back data (test set) should be predicted by the model. A measure of the error of prediction should be specified (RMSEP, $Q^2_{(ext)}$ or similar). A measure of similarity to the training set for each predicted observation is useful to indicate interpolation or extrapolation (such as a probability of membership) and should be given. The significance and confidence of regression coefficients should be reported to aid model interpretation.

6.6. Size and Selection of Test Set

In cases where no prediction set is available after model building, a training test set must be chosen from the data available. The number of observations will influence the choice of size for the test set. When dealing with many observations, a stringent test for the model is to remove a large proportion of the observations as a test set. This may be as much as a third to a half of the observations. Classical “leave one out” procedures are misleading when dealing with large numbers of observations as indicated earlier.

In non-ideal cases with low numbers of observations, test set prediction must still be carried out. In the case of $n < 10$ a “leave one out” method will be appropriate as each observation represents 10% of the data set. Each observation must be left out in turn and the prediction error or classification error calculated in each case. These may then be averaged to give an overall model predictability. For $n < 20$ a “leave 2 out” validation is appropriate. As n approaches 30, more ambitious sized test sets may be used. In cases where regular sized chunks of data are used as test sets the process

may be repeated over all fractions and the method success averaged across training/test set combinations.

If bias in test set selection is to be avoided, a rational method for selection of test sets should be used. Possible solutions to this include a random selection of the total dataset or use of a statistical method such as Design of Experiments using principal component scores [4]. This is a method for obtaining representative data free of bias from large multivariate datasets.

Only if the dataset is homogeneous with no clustering, should the following methods for calibrations be used - ordering of dataset by Y variable followed by selection of alternate observations (even/odd) or ordering of dataset by Y variable followed by selection of every third observation. There is a great risk with these latter two methods if groups or clusters are encountered within the data, which is a common situation in metabolic analyses. The risk is that by sorting the data, the X data will also be systematically sorted in terms of the clusters. The model will then reflect the clustering but may seriously mislead the investigator as the data may for instance relate to (say) chemical structure differences between groups and not the toxicity mechanism.

6.7. Biochemical interpretation of chemometric models

Wherever possible, data analysis methods should list the importance of variables and observations in the model and regression coefficients so that these may be interpreted for biological significance. Metabonomics provides ample opportunities for mis-interpretation of data. For example, in mammalian biofluid studies, the confounding effects of drug metabolites, animal handling, bacterial contamination all hamper the analysis. It is therefore recommended that the data analysis is done in close co-operation with the spectroscopists, chemists and biologists involved in the study in order that mis-interpretation is avoided.

It is important to realise that spurious "non-real" correlations may happen within data just by accident. The risk of spurious correlations increase with increasing number of variables but multivariate approaches (such as partial least squares, PLS) in combination with cross-validation and permutation validation minimise this risk. However, a "real" non-spurious correlation is still no guarantee that a cause and effect model will result.

Cause and effect models are only possible where the causes are systematically and independently varied (perturbed). However, non-causal real correlations are still very useful for classification, indirect calibration, diagnostics, and other applications in chemistry and biology. It is incorrect to interpret any model as causal, i.e., that "a change in these X-variables lead to this change in these Y-variables", but "this group of X-variables is correlated to this group of Y's", is valid.

7. The next steps

Chemometrics or multivariate data analysis is becoming more and more important as an integral part of scientific areas such as metabolic analyses. However, it is important to acknowledge that chemometrics will never replace the analytical chemist’s knowledge in interpreting data or the biologists/toxicologists interpretations of the interactions and mechanisms occurring in the studied biological systems.

Instead, chemometrics should be seen as a tool for creating robust and highly interpretable multivariate models with the aim of finding and understanding patterns and trends among observations and variables in large and complex data sets, as generated within metabonomics. These models will highlight important areas in the spectral and chromatographic data where further efforts on interpretation and metabolite identification should be put in. Based on this metabolic information the biological interpretation can then be carried out. With this chain of events in mind, it is easier to realise the importance of creating data sets that are suitable for chemometric or multivariate analysis by e.g. alignment, normalisation and binning. In the case of GC- and LC-MS based metabonomics these are all important issues to address in order to be able to compare, and extract information from, large sample populations in an organised and reliable fashion. The key point is that the original spectral or chromatographic resolution will always be retained in the raw data and that chemometrics is only a guide to find the relevant parts of the data to consider for interpretation.

This document represents a set of draft recommendations and standards and it would be advantageous and welcomed if a wider discussion now took place. Thus the next steps in this process are –

- To post this document to the SMRS web site to make it generally available.
- To circulate this document to all those on the SMRS e-mail contact list since this includes all those who helped prepare this paper and also others who have expressed an interest in the work.
- To invite, both via the web site and the email list, anyone interested in attending a discussion meeting in the USA on this subject to respond so that an estimate can be made of potential numbers of attendees.
- To abridge this document into a form suitable for publication in an appropriate journal.
- To construct a letter to send to journal editors for their feedback.

References

1. J.K. Nicholson, J. Connelly, J.C. Lindon and E. Holmes. Metabonomics: A platform for studying drug toxicity and gene function. Nature Rev. Drug Disc., l, 153-161 (2002).
2. O. Fiehn. Metabolomics - the link between genotypes and phenotypes. Plant Mol. Biol., 48**,** 155-171 (2002).
3. A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C.P. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo and M. Vingron. Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. Nature Genetics 29, 365-371 (2001).
4. J. Gullberg, P. Jonsson, A.Nordström, M. Sjöström, and T. Moritz. Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of Arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry. Anal. Biochem., in press (2004).
5. General Criteria of Competence for Calibration and Testing Laboratories. NAMAS Accreditation Standard M10, Edition 1, March 1989.
6. Good Laboratory Practice – The United Kingdom Compliance Programme. UK Department of Health, 1989.
7. J.L. Markley, A. Bax, Y. Arata, C.W. Hilbers, R. Kaptein, B.D. Skes, P.E. Wright and K. Wüthrich. Recommendations for the presentation of NMR structures of proteins and nucleic acids. J. Mol. Biol., 280, 933-952 (1998).
8. A. N. Davies, J. Lampert, R.L. Lancashire, P. Lampen, W. Conover, M. Frey, M. Grzonka, E. Williams, and D. Meinhart. Guidelines for the representation of pulse sequences for solution-state nuclear magnetic resonance spectroscopy. Pure Appl. Chem., 73, 1749-1764 (2001).
9. L. Wilkins. Guidelines of nuclear magnetic resonance computerized databases. Pure Appl. Chem., 67, 593-596 (995).
10. R.K. Harris, J. Kowalewski, and S.C. de Menezes. Parameters and symbols for use in nuclear magnetic resonance. Pure Appl. Chem., 69, 2489-2495 (1997).
11. R.K. Harris, E.D. Becker, S.C. de Menezes, R. Goodfellow, and P. Granger. NMR nomenclature. Nuclear spin properties and conventions for chemical shifts. Pure Appl. Chem., 73, 1795-1818 (2001).
12. S. Wold and O.H.J. Christie. Extraction of mass spectral information by a combination of autocorrelation and principal components models. Anal. Chim. Acta, 165, 51-59 (1984).
13. E. Johansson, S. Wold and K. Sjödin. Minimizing effects of closure on analytical data. Anal. Chem., 56, 1685-1688 (1984).
14. P. Jonsson, J. Gullberg, A. Nordstrom, M. Kusano, M. Kowalczyk, M. Sjostrom and T. Moritz. A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. Anal Chem., 76, 1738-45 (2004).

APPENDIX 1
Contact details for the SMRS Group contributors to this document

Name	Organisation	E-mail
Dr Alexander Amberg	Aventis	Alexander.Amberg@aventis.com
Dr Henrik Antti	University of Umea	henrik.antti@chem.umu.se
Dr Dorrit Baunsgaard	NovoNordisk	doba@novonordisk.com
Dr Richard Beger	NCTR (FDA)	RBeger@nctr.fda.gov
Dr Steve Bruce	Imperial College	s.bruce@imperial.ac.uk
Dr Glenn Cantor	BMS	glenn.cantor@bms.com
Dr Andrew Craig	Imperial College	a.craig@imperial.ac.uk
Dr Clare Daykin	University of Nottingham	clare.daykin@nottingham.ac.uk
Dr Mark Earll	Umetrics	mark.earll@umetrics.co.uk
Dr Nigel Hardy	University of Wales	nwh@aber.ac.uk
Dr John Haselden	GSK	john_n_haselden@gsk.com
Dr Elaine Holmes	Imperial College	elaine.holmes@imperial.ac.uk
Dr Hector Keun	Imperial College	h.keun@imperial.ac.uk
Dr Par Jonsson	University of Umea	Par.jonsson@chem.umu.se
Dr Erik Johansson	Umetrics	Erik.johansson@umetrics.com
Dr Sunil Kochhar	Nestle	sunil.kochhar@rdls.nestle.com
Dr Kerstin Kramer	Lilly	Kramer_Kerstin@lilly.com
Dr Arno Krotzky	Metanomics	Arno.Krotzky@metanomics.de
Dr Lois Lehman-McKeeman	BMS	lois.lehman-mckeeman@bms.com
Dr Johann Lindberg	AstraZeneca	Johann.Lindberg@astrazeneca.com
Prof John Lindon	Imperial College	j.lindon@imperial.ac.uk
Dr Mahendra Navarange	Medical Research Council	mahendra.navarange@imperial.ac.uk
Prof Jeremy Nicholson	Imperial College	j.nicholson@imperial.ac.uk
Mr Jake Pearce	Imperial College	jake.pearce@imperial.ac.uk
Dr Rob Plumb	Waters	rob_plumb@waters.com
Dr Jonathan Powell	Unilever	Jonathan.Powell@unilever.com
Dr Mike Reily	Pfizer	michael.reily@pfizer.com
Dr Don Robertson	Pfizer	donald.robertson@pfizer.com
Dr Susanna-Assunta Sansone	European Bioinformatics Institute	sansone@ebi.ac.uk
Dr Hartmut Schäfer	Bruker	hartmut.schaefer@bruker-biospin.de
Dr Ina Schuppe-Koistinen	AstraZeneca	Ina.Schuppe-Koistinen@astrazeneca.com
Dr Hans Senn	Roche	hans.senn@roche.com
Dr Manfred Spraul	Bruker	manfred.spraul@bruker.de
Dr Craig Thomas	Lilly	THOMAS_CRAIG_E@lilly.com
Dr Johan Trygg	University of Umeå	johan.trygg@chem.umu.se
Dr Frans Van der Ouderaa	Unilever	Frans.Van-der-Ouderaa@Unilever.com
Dr Elwin Verheij	TNO Pharma	verheij@voeding.tno.nl
Prof Ian Wilson	AstraZeneca	Ian.Wilson@astrazeneca.com
Dr Mark Viant	University of Birmingham	M.Viant@bham.ac.uk
Prof Svante Wold	Umetrics	Svante.wold@umetrics.com

APPENDIX 2

Notes and action points from the initial SMRS Group meetings

The initial meeting discussed the current perceived need for standardisation of metabolic analysis reporting and recognised that this would fulfil a need that is present now, but also for the future, even though some of the technologies used in metabolic analysis are not yet as mature as others. However, it was recognised that the approach has a number of differences and additional complexities compared to genomics. A finite number of variables are measured in genomics (according to the type of micro-array used), but in metabolic analysis there is an unknown number of possible variables depending on the type and sensitivity of the analytical method used and additionally a variety of analytical technologies can be employed, each with very different sensitivities, molecular information and data structures.

It was agreed that the research community was moving towards more openness and accessibility of data and it was concluded that reporting methods will have to encompass such diversity and to allow openness, and scrutiny of experimental methods, independent analysis of data sets by other groups will be necessary, and that standardisation of metabolic analysis information should harmonise with the analogous MIAME initiative in genomics [3]. A minimum requirement would be to describe what was done in a given study, for sample preparation, analytical measurements and data analysis, which analytical and chemometric technologies were used and details of any formal validation of data sets.

The standardisation would facilitate the rapid inclusion of data into public metabolic data bases (public or otherwise), for academic uses such as submission of supplementary data to journals to support publications, for internal use in pharmaceutical and other companies, and as an aid to submissions by pharmaceutical companies to regulatory bodies such as the FDA. It was pointed out that the FDA policy on new initiatives is essentially sponsor-driven through collaboration with all interested parties and that it will be desirable that any new types of data be compliant with an existing data structure for conventional meta-data known as SEND, developed by PharmQuest [www.pharmquest.com/send_consortium/overview.htm], who have signed an R&D agreement with the FDA for development of analysis tools specifically for toxicity data. The first model is now available and the FDA has issued draft guidelines for genomic submissions (these would have to be MIAME compliant). It was generally agreed that for non-clinical data (e.g. animal, sample details, clinical chemistry results, etc), the SEND format and categories should be used and that other data, e.g. metabonomics spectral data should be SEND compliant - a possible format for spectroscopy being JCAMP-DX [www.jcamp.org]. A discussion ensued on exactly what data should be included, e.g. raw spectra, processed spectra (e.g. filtered), binned and integrated regions of spectra as spreadsheet values, principal components scores and loadings, supervised chemometrics models, etc.

It was felt that for publications, data submission should not be compulsory but any submission would be to facilitate reviewing of manuscripts, that data held by companies would remain confidential until they decided to release it (this is current practice in other areas, e.g. protein structures) and that partial data set submissions are possible which would be different for different uses; only such data as was necessary to validate the conclusions drawn need be submitted. It was agreed that requirements

for journal and regulatory bodies could be different in that submission to journals places the data into the public domain, whilst submissions to regulatory bodies are confidential. One major area of concern related to intellectual property issues, and these would need to need to be addressed, particularly in terms of data ownership

It was agreed to hold a second meeting in March/April 2004 to draw up a more detailed proposal, bearing in mind that the initiative had been driven by toxicological applications but that the ultimate approach should include clinical and other human applications, as well as those in plant, microbiology, environmental or forensic sciences. It was planned to publish a summary of what was presented. Following a second meeting and drafting of a policy document, it was thought useful to hold a subsequent open formal meeting.. This meeting should have as wide a spread of expertise available to it as possible and possibly be held near Washington DC USA, so that informal representation by the FDA is possible

The background to the group formation and the subject matter covered in the first meeting was reprised in the second meeting. Other initiatives in related "omics" areas were summarised including the activities of the MGED Society, its Toxicogenomics Working group (TWG) and their collaboration with inter-related efforts, such us the I3C/HL7/CDISC Pharmacogenomics Standards Group and the Proteomics Standardization Initiative (PSI) within the Human Proteome Organisation (HUPO). An update on the SEND initiative was also provided (see Appendix 3 for more details).

The complexity of metabolic descriptions is more complicated than those for transcriptomics, having to take into account effects of other genomes such as from gut microflora, and their symbiosis with humans and animals. It was suggested that all methods are validated and that all descriptors and qualifiers should be validated and reported to regulatory/GLP standards. Awareness should be given to the different scales of analytical and biological variation. Other problems were discussed such being able to describe all data processing steps, especially if algorithms and methods are proprietary.

The several levels of reporting that would be appropriate for different applications were summarised in increasing order of completeness, public databases, submissions to support publication review and to allow journal readers access to data and regulatory submissions by pharmaceutical and other companies

Formalization of the SMRS initiative was agreed to be a short-term goal. It was agreed that it should join the SEND initiative. It was also proposed that it could become a society like MGED (transcriptomics) and PSI (proteomics) organised with a board of directors and an advisory board composed of academics, government personnel and industrial representatives and journal editors. The questions of who would build any databases and how this could be done were raised. Publicity for the group efforts was also needed.

The group agreed that the next step towards formal standard setting was to prepare a draft policy document and to submit an abridged form to a journal for publication. Additionally, it might be feasible to develop a publication on analytical validation for all "omics" for a journal such as Analytical Chemistry. Finally, a letter to journal

editors was considered. A precedent exists for this in the transcriptomics community in that the MGED Society has sent such a letter to a wide selection of journal editors proposing a paradigm for publishing transcriptomics data [www.separationsnow.com/basehtml/SepH/1,1353,6-1-1-0-0-news_detail-0-1440,00.html].

The full discussion document and other relevant documents will be posted to a new web page devoted to the SMRS group, registered and hosted from Imperial College, (www.smrsgroup.org) and an e-mail circulation list has been developed to keep all interested parties informed. Other interested parties wishing to participate in the metabolic standardisation effort are encouraged to contact the webmaster at the web address indicated above requesting inclusion in any future communications.

APPENDIX 3
A summary of standardisation initiatives for other omics approaches

Microarray Gene Expression Data (MGED) Society
The MGED Society [www.mged.org] is an international organisation of biologists, computer scientists, and data analysts, including representatives from academia, government and industry. Its goals are to establish standards for microarray data annotation and exchange, to facilitate the creation of microarray databases and the related software implementing these standards, and to promote the sharing of high quality, well-annotated data within the life sciences community. Future plans include the extension to other high throughput technologies, in close collaboration with other initiatives, such us the HUPO-PSI in the proteomics field (see below). The MGED Society has an established infrastructure for community outreach, including a central website (with links to working groups, open mailing lists and distribution of open source standards and software) and yearly international conferences (with tutorials, workshops and jamborees). The MGED Society Working Groups have developed a set of open source 'standards' which include:

- The Minimum Information About Microarray Experiments (MIAME) is a document which defines the minimum descriptors that must be reported about array based gene expression monitoring experiments, in order to ensure the interpretability of the results, as well as potential verification by third parties. MIAME provides a framework for capturing information and is a standard for data content not a format standard. The MIAME checklist [A3.1] is a condensed description of MIAME principles, designed to help authors, reviewers and editors of scientific journals to meet the requirements and aimed to make data supporting publications publicly available. As the array technology evolves, MIAME will be extended to reflect this and new concepts will be added.

- The MicroArray Gene Expression (MAGE) standards [A3.2] are the MAGE Object Model (MAGE-OM) and MAGE Markup Language (MAGE-ML). MAGE provides the formal standard that specifies the communication protocols, ensuring software interoperability and encoding all MIAME required information. MAGE is a stable, formal specification of the Object Management Group (OMG) (Gene Expression, v1.1) [www.omg.org/technology/documents/formal/gene_expression.htm]. The current version has been designed to model microarray experimental protocols and technologies. However, the MAGE Working Group is working with the HUPO-PSI group towards a model to support multiple functional genomics experimental domains.

- The MGED Ontology [A3.3] provides standard terms for the annotation of microarray experiments by investigators, and for software and database developers. These terms allow structured queries of elements of any experiment and allow unambiguous description of how the experiment was performed. A core MGED Ontology has been established and will remain unchanged to facilitate software development. However, as the MGED

Society extends its mission to other high throughput techniques in functional genomics, a second layer of ontology will be built. Currently, the MGED Ontology working group is collaborating with the HUPO-PSI group (see below) to add descriptive terms for proteomics.

The response from the scientific community to these 'standards' has been extremely positive [A3.4]. Currently most of the major scientific journals and some funding agencies require publications describing microarray experiments to comply with MIAME. Furthermore, the adoption and implementation of the MGED 'standards' by public [A3.5-A3.7, www.niehs.nih.gov/nct] and institutional [A3.8-A3.10] databases and several microarray informatics tools [www.rosettabio.com/products/resolver/default.htm; www.silicongenetics.com; www.bioconductor.org] has greatly improved the accessibility of the datasets.

The Toxicogenomics Working Group (TWG) has been established recently under the MGED umbrella, as part of a collaborative undertaking with the ILSI Health and Environmental Sciences Institute's (HESI) 'Technical Committee on the Application of Genomics to Mechanism Based Risk Assessment' [hesi.ilsi.org/publications/pubslist.cfm?pubentityid=120] the EMBL-EBI [www.ebi.ac.uk/microarray/Projects/tox-nutri], NIEHS NCT [www.niehs.nih.gov/nct] and NCTR-FDA [www.fda.gov/nctr]. The MGED TWG has initiated the extension of MIAME as applied to array-based toxicogenomic experiments. The MIAME/Tox checklist [MIAME/Tox checklist: http://www.mged.org/Workgroups/tox/tox.html] extends the MIAME checklist to provide a structured annotation and framework for capturing information associated with the in-life component of toxicogenomic experiments and other MIAME-based checklists are being developed by environmental genomics and nutrigenomics communities (MIAME/Env and MIAME/Nut) [envgen.nox.ac.uk/miame/miame_env.html; www.nugo.org], to fulfil the specific requirements of their array-based applications. The new MGED TWG group will include other communities interested in standardisation such nutrigenomics and environmental genomics. To reflect this expansion, the new group will also change its name to the Reporting Structure for Biological Investigations (RSBI) Working Group [www.mged.org/Workgroups/rsbi]. Currently the RSBI Working Group is developing a proposal to re-structure these MIAME-based checklists into a tiered structure concept, a modular context dependent structure, splitting the design of the investigations and the sample description from both the requirements of a specific application (toxicology, environment and nutrition) and the -omics technologies used. Similar to MIAME, this reporting structure aims to identify the minimal descriptors of an investigation and organise the information in a structured manner to promote harmonisation across high throughput technologies and their domains of application (where a clear overlap exists), but retaining the specifics of each discipline. Ultimately this may assist in the identification of steps or processes that can be considered quality metrics to validate of technology used and/or as checkpoints for performance assessment. Members of the RSBI Working Group have strong interactions with other efforts, including PSI, the Pharmacogenomics Standards Group and ECVAM (see below).

The Proteomics Standards Initiative (PSI)

PSI [http://psidev.sourceforge.net] is an initiative started within the Human Proteome Organization (HUPO), including representatives from the main protein databases,

government and industry. The goal is to define community standards for data representation in proteomics to facilitate data comparison, exchange and verification. Currently PSI is developing standards for two key areas of proteomics, mass spectrometry and protein-protein interactions data, as well as a standardised general proteomics format. A set of open source 'standards' are being developed which include:

- An XML standard data exchange for Protein-Protein Interaction [A3.11] to allow the several well-established databases to exchange data. Within mass spectrometry, PSI is developing a standard representation of experimental spectra in the context of the experimental setup. Lastly, a standardised format for proteomics data is being developed. This is able to accommodate data from different experimental methods, while sharing common components like sample description and description of the relevant proteins.

- The Minimum Information About Proteomics Experiments (MIAPE) [http://psidev.sourceforge.net/gps/misc/MIAPE_Principles.doc], a checklist defining the minimum descriptors that must be reported in a paper or a submission to a database, is a work in progress. Description of biological systems in MIAPE is key, and will be developed collaboratively with MGED members to avoid duplication of effort.

- Whilst the PSI-MS Working Group is working with the American Society for Mass Spectrometry (ASMS), an ontology of clearly defined general proteomics terms is being developed by the PSI-GPF Working Group for the construction of unambiguously worded data files. To avoid duplication of effort this group is working closely with MGED Ontology Working Group.

Standard for Exchange of Non-clinical Data (SEND) Consortium

SEND, as mentioned above, is a consortium formed among the pharmaceutical industry, contract laboratories, software developers, and the FDA. The goal is to develop a common model to define metadata for all aspects of a typical animal toxicity study, such as body weights, food consumption, organ weights, clinical pathology, macroscopic/microscopic findings, reproductive toxicity parameters, and tumour findings, among others. The focus is to facilitate electronic data submission to the FDA. An open source 'standard' has been developed. The SEND model v1.4 provides a general framework for submitting datasets structured as a flat file with pre-defined categories. A Cooperative Research and Development Agreement (CRADA) has been established between the FDA Center For Drug Evaluation and Research (CDER) and PharmQuest to develop and evaluate software tools for receiving, storing, viewing and analyzing nonclinical data based on SEND model. SEND has been modeled after CDISC's Submission Data Standards (SDS) model v.3 (see below) and plans are to merge and harmonise both models to form a new standard for data presentation – the Study Data Tabulation Model (SDTM).

Clinical Data Interchange Standards Consortium (CDISC)

CDISC is an open, multidisciplinary, non-profit organization committed to the development of worldwide biopharmaceutical industry standards, vendor-neutral, platform-independent data models to support the electronic acquisition, exchange,

submission and archiving of clinical trials data and meta-data [www.cdisc.org]. The focus is to streamline regulatory electronic submissions with a consistent framework and format. An open source 'standard' has been developed. CDISC Submission SDS model v3.1 consists of two documents: the SDTM, which represents the underlying conceptual model behind the SDS standards, and the CDISC V3.1 SDTM Implementation Guide (SDTM-IG), which includes the detailed domain descriptions, assumptions, and examples. SDTM will help automate the largely paper-based clinical trials research process and foster easier communication and collaboration among clinical researchers. The SDTM has been implemented first for human drugs but later will become the single model for both clinical and non-clinical data with additional variables added as needed. A CRADA has been established between the FDA CDER and Lincoln Technologies and to develop and evaluate software tools for receiving, storing, viewing and analyzing non-clinical data based on CDISC's SDS models. An additional CRADA has been established with IBM for the development of Janus Data Warehouse, designed to integrate with SDTM and associated software tools, but not limited to them. The Janus Data Warehouse is intended to capture data submitted to the FDA collected from clinical trials (and animal toxicity studies) along with a description of study protocols to permit a high degree of automated analysis.

Pharmacogenomics Standards Group

This is a joint project of CDISC, Health Level Seven (HL7) [www.hl7.org] and Interoperable Informatics Infrastructure Consortium (I3C) [www.i3c.org], including industry, software developers and governmental representatives. CDISC has formal liaison to HL7, through the Regulated Clinical Research Information Management (RCRIM) Committee, co-chaired by HL7, CDISC and the FDA. Through this relationship, CDISC has been working to harmonise its clinical trial data models with the HL7 Reference Information Model (RIM) in the context of healthcare information. HL7 is an ANSI-accredited standards developing organization that provides standards for the exchange, management and integration of data that supports clinical patient care and the management, delivery and evaluation of healthcare services. I3C is a discovery informatics consortium, which aims to eliminate barriers to application interoperability, data integration and knowledge flow. The Pharmacogenomics Standards Group was formed in November 2003 at a workshop organized the by the Drug Information Association (DIA), FDA, Pharmacogenetics Working Group (PWG), Pharmaceutical Research and Manufacturers Of America(PhRMA) and Biotechnology Industry Organization (BIO) to review the FDA draft "Guidance for Industry - Pharmacogenomic Data Submissions". The goal of this joint project is to define the requirements for pharmacogenomics submission to FDA and clarify data formats, standards. This project is focusing on the use of pharmacogenomics and toxicogenomics data to support pharmacological conclusions, and thus encompasses both a toxicology and a micorarray component. Currently, three sub-teams have been formed, based upon the DIA Pharmacogenomics Standards Workshop, namely: pre-clinical/non-clinical genomics, clinical pharmacology and clinical genomics The sub-teams meet regularly by phone conference and in joint meetings. There is a consensus within this group to use existing standards (e.g. MIAME, MAGE, SEND, CDISC) extend them if needed and only create new standards where necessary.

The European Centre for the Validation of Alternative Methods (ECVAM)

ECVAM [http://ecvam.jrc.cec.eu.int/index.htm] seeks to promote the scientific and regulatory acceptance of alternative methods that could reduce, refine or replace the use of laboratory animals. ECVAM is working with the US Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) and National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) [http://iccvam.niehs.nih.gov/home.htm] to investigate the specific considerations necessary for adequate validation of array-based toxicogenomics-based test methods. A description of biological systems is a critical part of such a test. The current focus in on establishing a process that will facilitate future regulatory acceptance of scientifically valid toxicogenomics-based test methods. The possibility of transferring validation processes and principles for toxicogenomics to other new technologies (e.g. proteomics, metabolic analysis) is being also discussed. Recommendations are being developed to cover topics such as biological systems, methodological/technical issues, data analysis, and data storage with a view towards regulatory acceptance.

Other efforts

Other organizations and committees are also tackling data standardisation issues in a wider context of using emerging high throughput approaches. The International Program on Chemical Safety (IPCS) [http://www.who.int/ipcs/en/] is a joint programme of three cooperating organisations - International Labour Organization, United Nations Environment Network and World Health Organization - implementing activities related to chemical safety. In collaboration with Organisation for Economic Co-operation and Development (OECD) [http://www.oecd.org/home/] IPCS organizes a series of workshops to explore the future international activities related to the use of toxicogenomic methods in chemical assessment.

The National Academy of Sciences (NAS) Committee on Emerging Issues and Data on Environmental Contaminants [http://dels.nas.edu/emergingissues/index.asp] is a public forum for communication among government, industry, environmental groups, and the academic community about emerging evidence and issues in toxicogenomics, environmental toxicology, risk assessment and exposure assessment. The Committee will develop a framework for how the emerging field of genomics will be incorporated into risk assessment.

The Distributed Structure-Searchable Toxicity (DSSTox) Database Network is a project by the US EPA [http://www.epa.gov/nheerl/dsstox/], providing a community forum for publishing standard format, structure-annotated chemical toxicity data files for open public access.

A3.1. Ball C.A., Sherlock G., Parkinson H., Rocca-Serra P., Brooksbank C., Causton H.C., Cavalieri D., Gaasterland T., Hingamp P., Holstege F., Ringwald M., Spellman P., Stoeckert C.J. Jr, Stewart J.E., Taylor R., Brazma A., Quackenbush J. An open letter to the scientific journals published in: *Science* 2002 298, 539; *Bioinformatics* 2002 18, 1409; *The Lancet* 2002 360, 1019.

A3.2. Spellman P.T., Miller M., Stewart J., Troup C., Sarkans U., Chervitz S., Bernhart D., Sherlock G., Ball C., Lepage M., Swiatek M. Marks W.L., Goncalves J., Markel S., Iordan D., Shojatalab M., Pizarro A., White J., Hubley R., Deutsch E., Senger M., Aronow B.J., Robinson A., Bassett D., Stoeckert C.J. Jr., Brazma A.. Design and implementation of microarray gene

expression mark-up language (MAGE-ML). *Genome Biol.* 3, Issue 9, research0046 (2002). [Open access electronic journal].

A3.3. Stoeckert C.J., Causton H.C., Ball C.A. Microarray databases: standards and ontologies. *Nature Gen Suppl S*, 2002 32, 469-473. Microarray standards at last. *Nature* 2002 419, 323.

A3.4. Edgar R., Domrachev M., Lash A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002 30, 207-10.

A3.5. Ikeo K., Ishi-i J., Tamura T., Gojobori T., Tateno Y. CIBEX: center for information biology gene expression database. *C R Biol*. 2003 326, 1079-82.

A3.6. Brazma A., Parkinson H., Sarkans U., Shojatalab M., Vilo J., Abeygunawardena N., Holloway E., Kapushesky M., Kemmeren P., Lara G.G., Oezcimen A., Rocca-Serra P., Sansone S.A.ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*. 2003 31, 68-71.

A3.7 Waters M.., Boorman G., Bushel P., Cunningham M., Irwin R., Merrick A., Olden K., Paules R., Selkirk J., Stasiewicz S., Weis B., Van Houten B., Walker N., Tennant R. Systems toxicology and the chemical effects in biological systems knowledge base. Environ. Health Perspect. *Toxicogenomics* 2003 111, 811-824. Chemical Effects in Biological Systems. (CEBS) Knowledge Base at NIEHS NCT.

A3.8. Gollub, J., Ball, C.A., Binkley, G., Demeter, J., Finkelstein, D.B., Hebert, J.M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J.C., Schroeder, M., Brown, P.O., Botstein, D., Sherlock, G. The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res* 2003 31, 94-96.

A3.9. Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., Quackenbush. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 2003 34, 374-378.

A3.10. Tong W., Cao X., Harris S., Sun H., Fang H., Fuscoe J., Harris A., Hong H., Xie Q., Perkins R., Shi L., Casciano D. ArrayTrack--supporting toxicogenomic research at the U.S. Food and Drug Administration National Center for Toxicological Research. *Environ Health Perspect Toxicogenomics* 2003 111, 1819-26.

A3.11 Hermjakob H., Montecchi-Palazzi L., Bader G., Wojcik R., Salwinski L., Ceol A., Moore S., Orchard S., Sarkans U., von Mering C., Roechert B., Poux S., Jung E., Mersch H., Kersey P., Lappe M., Li Y.X., Zeng R., Rana D., Nikolski M., Husi H., Brun C., Shanker K., Grant S.G.N., Sander C., Bork P., Zhu W.M., Pandey A., Brazma A., Jacq B., Vidal M., Sherman D., Legrain P., Cesareni G., Xenarios L., Eisenberg D., Steipe B., Hogue C., Apweiler R. The HUPO PSI Molecular Interaction Format - A community standard for the representation of protein interaction data. *Nature Biotech.* 2004 22, 177-183.

APPENDIX 4
Typical sample handling procedures for NMR spectroscopy

1. Biofluids
Rat and mouse urine samples can be prepared for NMR spectroscopy into covered 96 well plates using a Bruker SampleTrack system and a Gilson 215 Prep robot or equivalent technology.

In order to provide some stabilisation of the urinary pH, 400 µl of rat urine is mixed with 200 µl of buffer (a 0.2 M solution of Na_2HPO_4 in H_2O and a 0.2 M solution of NaH_2PO_4 in H_2O:D_2O (80:20) are prepared. TSP is added at 1 mM concentration and sodium azide at 3 mM) in 96 well plates using a Bruker, or similar, preparation robot. In the case of the mouse, 200 µl of urine is mixed with 200 µl H_2O and 200 µl buffer. For plasma or serum, aliquots (200 µl) are added to 400 µl of saline (0.9% NaCl in water/D_2O, 90:10, no TSP). These volumes can be adjusted according to the NMR probe used. The first and last sample in each plate is a blank where H_2O is substituted for urine. This provides a check for carry-over between samples and prevents urine remaining in the probe at the end of the run. The well plate is centrifuged at 4,000 rpm for 5 min to remove insoluble material, prior to positioning in the NMR flow injection Gilson robot.

500 µl of sample is injected into the probe with no air gaps in between sample and push solvent. The push solvent is 0.1 % sodium azide in H_2O. After each run the system is flushed using hydrogen peroxide (10%), followed by HCl (0.1M) and finally H_2O containing azide (0.1%). Once a week the probe is flushed manually with disinfectant and the solution allowed to stand in the probe for 1 h. Thorough cleaning of the probe by flushing for several hours after removing from the magnet will also be carried out periodically as required by continuous monitoring of performance.

2. Preparation of tissue extracts for NMR studies
(contributed by Mark Viant)
Introduction
Although biofluids are in general easier to collect than tissue samples or biopsies, and facilitate time-course studies of changes in metabolism within individual organisms, the metabolic information obtained from a biofluid such as plasma, however, is effectively an integration of the individual metabolic changes occurring within each of the animal's organs. In order to determine organ-specific metabolic fingerprints, which are of interest when investigating certain diseases or sites of toxicity, or when the metabolic interaction between different tissues is of interest, studies of tissues becomes highly desirable.

It is important to rapidly collect and freeze the tissue sample, so as to immediately 'quench' metabolism and preserve the metabolite concentrations. Samples can then be stored at -80ºC to temporarily prevent any metabolic decay. A method is needed for mechanically disrupting the tissue to allow the extraction of the low molecular weight metabolites. This is to de-proteinise the sample to permanently halt metabolism, and to extract only those metabolites of interest (e.g., lipids, carbohydrates, amino acids and other small metabolites) from the tissue while leaving other compounds in the tissue pellet (e.g., DNA, RNA, proteins). Unlike mass spectrometry, NMR is moderately tolerant of salt within samples and hence its removal is not required.

Finally, it is necessary to optimise the solution for high resolution NMR spectroscopy by re-suspending the metabolite extract in an appropriate deuterated solvent buffer and adding an NMR internal standard.

Method overview
The ideal sample size is 100-150 mg (wet mass) although as little as 20 mg can be used. A manual grinding method with a mortar and pestle is ideal to mechanically disrupt tissue for both 'hard' and fibrous (e.g., muscle) or 'soft' (e.g., liver or brain) tissue types. Several extraction methods can be used, the choice of which depends upon the polarity of the metabolites that are required. Perchloric acid can be used for the extraction of polar metabolites only. If both polar and lipophilic metabolites are desired then a methanol/chloroform procedure to yield two separate fractions can be used. Prior to NMR analysis, all polar tissue extracts can be prepared as described below.

Extraction of polar metabolites using perchloric acid
1. Grind the frozen tissue in a pre-cooled mortar, and transfer the frozen powder to an eppendorf tube sitting in dry ice. Weigh the eppendorf tube to calculate the wet tissue mass of the sample.
2. Add 5 mL/g of 6% ice-cold perchloric acid, vortex, then place on ice for 10 min.
3. Centrifuge (12,000g, 10 min, 4°C).
4. Transfer a known volume (e.g., 500 μL) of supernatant to a clean eppendorf tube, and then neutralize to pH 7.4 with 2M K_2CO_3 using a previously established ratio (typically 210 μL K_2CO_3 per mL supernatant). Incubate on ice for 30 mins to precipitate the potassium perchlorate salts.
5. Check that the pH of each sample is 7.4 (± 0.2 units).
6. Centrifuge (12,000g, 10 min, 4°C).
7. Transfer a known volume (e.g., 500 μL) of supernatant to an eppendorf tube, freeze in dry ice, and then freeze-dry the tissue extract overnight. Store at -80°C until required.

Combined extraction of polar and lipophilic metabolites using methanol/chloroform
1. Follow the same procedure as above for grinding and weighing the frozen tissue, except collect the powdered sample in a glass vial.
2. Add 4 mL/g methanol and 0.85 mL/g water to the sample and vortex.
3. Add 2 mL/g chloroform to the sample and vortex.
4. Mix the sample, on ice, using an orbital shaker at 300 rpm for 10 min. The solution will be mono-phasic.
5. Add 2 mL/g chloroform and 2 mL/g water to the sample and vortex.
6. Centrifuge (1,000g, 15 min, 4°C). The solutions will separate into an upper methanol/water phase (with polar metabolites) and a lower chloroform phase (with lipophilic compounds), separated by protein and cellular debris.
7. Transfer the upper and then lower layers of each sample into separate glass vials.
8. Remove the solvents from the samples using a speed vacuum concentrator and then store at -80 °C until required.

Preparation of polar extracts for NMR spectroscopy

1. Re-suspend the polar tissue extracts in 550 μL NMR buffer (100 mM sodium phosphate buffer, pH 7.4, in D_2O, containing 0.5 mM sodium 3-(trimethylsilyl)proprionate-2,2,3,3-d_4 (TMSP)) and vortex.
2. Centrifuge (12,000g, 5 min) and then transfer 520 μL into an NMR tube.

Preparation of lipophilic extracts for NMR spectroscopy
1. Re-suspend the lipophilic tissue extracts in 550 μL deuterated NMR solvent (2:1 mixture of chloroform-d ($CDCl_3$) and methanol-d_4 (CD_3OD), containing 0.5 mM tetramethylsilane (TMS)) and vortex.
2. Centrifuge (1,000g, 5 min) and then transfer 520 μL into an NMR tube.

APPENDIX 5
Typical NMR data acquisition and pre-processing parameters

The ^{1}H NMR spectra of the urine can be measured using a specified water suppression pulse sequence such as noesypresat which, if a Bruker instrument is used, is called noesypr1d. This has the form –RD-90º-t-90º-t_m-90º-ACQ, where RD is a relaxation delay, t is a short delay typically of about 3 μs, 90º represents a 90º RF pulse, t_m is a mixing time and ACQ is the data acquisition period.

In Bruker terminology, the O1 value is set to the H_2O resonance, the saturation power is set near 53 dB but will be instrument dependent, t = 3 μs, t_m ~ 80-100 ms, the sample temperature is set at 300K, SW = 12019.2 Hz, TD = 32K, D8 = 100 ms, RD = 2.0 s, NS = 64, RG=128. This results in a total acquisition time of about 4 min per sample. For processing, LB = 1 Hz is applied and SI = 32k is used.

It is recognised that a target spectral resolution is required. This has been set at a line width at half height on the TSP resonance of 2.5 Hz and sufficient resolution in the "hump" test to ensure that the ^{29}Si satellite peaks are evident. For urine samples containing high levels of protein this will not be achieved. Thus failure to reach this resolution will not cause the data acquisition to be halted. A target signal-noise ratio on the TSP signal is not included at this stage although this might be useful for making comparisons of spectrometer performance at different field strengths and NMR probe volumes. A target signal-noise ratio on the peaks from the endogenous metabolites is not possible due to some toxins causing large changes in urinary volumes and hence dilution.

For serum or plasma, water-suppressed 1D ^{1}H NMR spectra can be measured using the noesypr1d pulse sequence. In addition, spin-echo spectra can be measured using the CPMG sequence (i.e. cpmgpr), J-resolved spectra (jrespr) and diffusion-edited spectra (ledbpgskspr). The NMR parameters for the noesypr pulse sequence (as for urines) are NS = 128, TD = 32K, D1 = 2s, SW = 12 KHz, AQ = 1.36 s, D8 = 100 ms.

The CPMG pulse sequence has the form –RD-90º-(t-180º-t)$_n$-ACQ, where the definitions are as above, plus 180º is a 180º RF pulse, t is the spin-echo delay and n represents the number of loops. Typically NS : 8 * n, DS : 16, NS = 128, TD = 32K, SW = 12 KHz, RD = 2 s, AQ = 1.36, n = 80, t = 400 μs, with a total echo time = 64 ms and irradiation at the water peak during RD.

The J-resolved pulse sequence has the form –RD-90º-t1-180º-t1-ACQ, where t1 is an incremented time period, the RD delay is 1-5 * T1, NS: 4 * n, DS: 16, NS = 16**,** TD = 32K (F2), 32 (F1) SW = 12 KHz (F2), 40 Hz (F1), AQ = 1.36, RD = 2 s. Although the 2-dimensional J-resolved spectrum is obtained, pattern recognition is often carried out on the sum or skyline F2 projection.

The diffusion edited pulse sequence using bipolar gradients and the LED scheme has the form -RD-90º-G_1-180º-G_1-90º-G_2-Δ-90º-G_1-180º-G_1-90º-G_2-τ-90º-acquire FID, where RD is a relaxation delay, 90º is a 90º RF pulse, G_1 is the pulsed field gradient that is applied to allow editing, 180º is a 180º RF pulse, G_2 is a spoiler gradient applied to remove unwanted magnetisation components, Δ is the diffusion delay during which the molecules are allowed to diffuse, and τ is a delay to allow the

longitudinal eddy currents caused within the sample to decay. RD is a relaxation delay of about 1-5 * T1, NS = 64, TD = 32K, W = 12 KHz, AQ = 1.36 s, RD = 2 s, D20 = 0.1 s, D16 = 0.1 ms, D21=5 ms, and for processing LB = 1 Hz is applied and SI = 32k is used.

An alternative approach, which allows the identification and quantitation of metabolites directly, involves using the software supplied by Chenomx Inc. Preparation of the samples requires only the addition of a chemical shift standard, such as DSS or TSP at a known concentration, and D_2O to a final concentration of 10%. Chenomx also recommends that the user add NaN_3 to prevent bacterial growth and a user may wish to add imidazole to aid in detecting the pH of the sample, but this is not necessary if the user knows the pH. It is further suggested that care be taken in acquiring data such that the line width and peak shape of the chemical shift standard be optimized and the use of the presat-noesy or noesypr1d pulse experiment with a 1 s recycle delay, 100 ms mixing time and 4 s acquisition time is recommended. Proton irradiation at the water resonance should be turned on during the recycle delay and the mixing time and a 12 ppm spectral width should be used to maximize resolution. To process, a 0.5 Hz line broadening and zero-filling to 128 or 256k should be applied. The Chenomx software and pH sensitive spectral databases may be used to deconvolve NMR spectra into metabolite concentrations.

APPENDIX 6
Some definitions used in Chemometrics

Training Set
The training set is the set of data used to build a model

Test Set (Also called External Test set)
The test set is a set of data NOT used to build the model. From the test set, it is possible to get a good estimate of the predictivity

Prediction set
The prediction set is a completely new set of data not available at the time of model building. From the prediction set, it is possible to obtain predictivity. When the training set is representative of the prediction set the internal predictivity estimate is expected to be similar to the predictivity estimate. Large discrepancies between such values indicate failure to select a representative and diverse training set.

Internal Cross Validation
Internal cross validation is a method used by some software packages to simulate the effect of predicting new data. While it is useful in model building as a guide to predictive ability it should not be relied upon entirely. A significant proportion of the dataset is removed and the model built on the data left behind. The left out data is predicted and the difference between the two computed. This is repeated across the whole data set (all folds) and summed into a measure of model predictability.

Another validation method for regression modelling is permutation testing, where the order of the Y variable is randomly scrambled. Models are built on many scrambled datasets and the fit and predictability are assessed for each. For a valid regression the scrambled models should be much worse than the original data. This method reliably detects spurious correlations.