

# The challenges of recording phenotype in a generalizable and computable form

PM Nadkarni

*Yale University School of Medicine, New Haven, CT, USA*

*The Pharmacogenomics Journal* (2003) 3, 8–10. doi:10.1038/sj.tbj.6500153

Pharmacogenomics research attempts to ‘correlate genotype with phenotype’ in the context of pharmacotherapy. Public databases of interest to pharmacogenomics researchers will therefore need to store phenotypic data. One goal of electronic storage will be to make it amenable to programs that perform ‘data mining’, a form of automated hypothesis discovery.<sup>1</sup>

Genotype is reasonably straightforward to represent computationally, as variations from a consensus sequence. Most of these variations are simple substitutions in the form of single nucleotide polymorphisms (SNPs), although one must also consider insertions, deletions and substitutions of greater length, such as variable numbers of tandem repeats (VNTRs). Much research today, rather than focusing on individual variants in isolation, considers the set of several such variants that are inherited as a unit (the haplotype). Existing public databases such as dbSNP,<sup>2</sup> which in turn point to databases containing consensus sequence such as Genbank, address the representation of both genotype as well as haplotype.<sup>3</sup> By contrast, representing phenotype—‘the outward, physical manifestation of internally coded, inheritable information’<sup>4</sup>—in a generic, computable way is much more problematic.

Below, we first consider the extreme variability of phenotypic parameters,

which are the source of the challenge. We describe how phenotype is represented in individual data sets. We then discuss how individual phenotypic parameters in a data set must be adequately described in order to be comprehensible and minimize the possibility of misinterpretation, and how even the nature of these descriptors depends on the broad category of phenotype that one is describing. We finally consider the requirements of software tools to assist submission of phenotypic data to an electronic repository, and the means by which such standardization of such software could be facilitated.

## CHALLENGES IN CHARACTERIZING PHENOTYPE

- Phenotype is not a single entity: it is a set of parameters. The universe of parameters constituting ‘phenotype’ is highly variable: function can be characterized at a molecular, organelle, cellular, organ-system or whole-organism level.
- The parameters are specific to the gene or genes that are being studied. The phenotypic parameters for breast cancer susceptibility are quite different from that of acute intermittent porphyria. Across all genes or genetic disorders of interest, the total number of parameters would range in the hundreds of thousands.
- Phenotype does not necessarily imply a single causative gene. The easy problems of classical pharmacogenetics, where a single mutation leads to fairly dramatic effects in

the right circumstances—for example, pseudocholinesterase deficiency manifesting as prolonged paralysis after succinylcholine administration—will become increasingly scarce. Multigenic diseases like diabetes, hypertension and obesity are difficult problems. For such diseases, as knowledge about causative mechanisms and regulatory pathways accumulates, the definition of phenotype becomes progressively more refined. In diabetes, for example, the original phenotypic definition of inability to utilize glucose (high blood sugar, glycosuria and a battery of clinical symptoms) was refined once insulin was discovered. We now had to distinguish between failure to produce enough insulin vs failure to respond to insulin. The latter category has since been broken down into numerous subcategories.<sup>5</sup>

A consequence of the continual refinement of phenotypic parameters is that data more than a few years old will rapidly diminish in value as a source for hypothesis discovery. Newly discovered parameters that come to be regarded as mandatory for evaluation of the disease of interest will be absent from that data set. The only way to obtain values for these parameters will be to approach the investigator who originated the data, and seek access to those patients in the dataset who are still alive. In many cases this will not be possible, either because the data needed are time-dependent (eg, drug levels) or because the institutional review board required subject anonymization or prohibited follow-up.

## REPRESENTING PHENOTYPIC DATA IN A DATA SET: METADATA

Phenotypic data typically exist as one or more ‘flat files’ consisting of columns, each of which represents a parameter, and rows, representing instances of experimental subjects, for example, patients or biospecimens. (For studies involving numerous para-

meters with long-term follow-up of subjects, a data set would consist of several such files—in effect, a relational database.) To be interpretable, a ‘data dictionary’—a document that describes in some depth what each column in each file represents—must accompany the data. A data dictionary is an example of ‘metadata’—data that describe data. *The challenge of standardizing the computational representation of phenotype then reduces to the problem of standardizing the metadata.* Such standardization efforts are taking place in other bioscience fields, such as the field of microarray experiments, where a data interchange standard called MAGE-ML (micro-array and gene expression markup language) is evolving.<sup>6</sup>

### Requirements of Phenotype Metadata Descriptors

Some of the columns in a phenotype data set represent information that identifies the subject or sample, for example, a subject study number or biospecimen ID, plus demographic variables such as age, sex and race. Although different researchers may encode variables such as sex and race in different ways—race also tends to be recorded at varying levels of granularity—the interpretation of such data is relatively straightforward. The experimental parameters that are specific to a study, however, are more challenging: even the metadata descriptors that fully describe such a parameter depend on the nature of the parameter.

Consider, for example, a parameter based on a lab test, which may either be a standard one or a newly devised technique. LOINC (logical observations, identifiers, names and codes)<sup>7,8</sup> is an existing standard devised for the purpose of interchange of data for lab values and vital signs. The full set of descriptors for a laboratory test parameter includes:

- The name of the parameter.
- A brief description, in cases where this parameter is not a standard one in routine use.
- The source of the biological sample: blood, plasma, serum, urine, etc.

- The timing of the sample: random vs collected at a particular time; single sample vs a cumulative sample collected, for example, over 24 h.
- The units of measurement.
- The lower and upper limit of ‘normal’ values, if this is known. Note that, in some cases, these can vary with age, sex and physiological state, for example, pregnancy and lactation. Even for ‘standard’ tests like total serum cholesterol, the ‘normal’ range is known to vary slightly from lab to lab.
- A bibliographic reference to the test method, if standard. If nonstandard, a reasonably detailed description of how it is performed must be provided—whether the test is a chemical or immunoassay, the reagents used, and so on.

For clinical (*in vivo*) measurements, one must describe the conditions under which the measurement was conducted. For example, a pair of systolic/diastolic blood pressure readings must be accompanied by details about the position of the subject (standing vs sitting vs supine), the instrument used (mercury vs. aneroid device—the latter is known to give highly variable readings if not periodically recalibrated), and which limb was used for recording.

For *in vitro* assays of enzymatic activity in biospecimens (eg, a microsomal enzyme activity assay), one might record:

- The chemicals/substrates used in the test.
- The concentration of the substrate (in case the kinetics is suspected to be nonlinear).
- The end-products of the enzymatic reaction.
- The conditions of the assay: nature of the buffer, the pH and temperature and so on.

Identifying the requisite descriptors for many types of molecular or biochemical phenotype data is more challenging, in part because the experimental areas themselves continue to evolve rapidly. For example:

- The same parameter may be studied by a variety of techniques. For

example, the polymerization of sickle hemoglobin (HbS), the first disease identified as caused by a molecular mutation, has been studied by physical techniques as varied as ultracentrifugation, viscosimetry, NMR and differential interference contrast (DIC) microscopy.<sup>9</sup> Each of these techniques requires its own set of descriptors.

- Measurement of associated gene expression levels through microarray technology is increasingly popular. These levels are influenced significantly in experimental animals by several confounding variables such as the time of the day the specimen was collected as well as whether the animal being studied was caged in isolation or along with other animals. Based on the system being studied, it may be important to describe the experimental conditions in enough detail to make it clear that known confounding variables were considered.

### REQUIREMENTS OF SOFTWARE TOOLS FOR PHENOTYPE DATA SUBMISSIONS

It is clear that having to supply such detail for every single parameter in a phenotypic data set can be a highly onerous task for data submitters. One must note, however, that publication in scientific journal generally mandates such details, and for electronic publication standards cannot be dramatically less stringent. Nonetheless, if public databases are to encourage submission of phenotypic data, the designers of such databases must strive to reduce the amount of manual labor required of submitters. Software tools accompanying electronic phenotypic repositories must provide the equivalent of Genbank’s SEQUIN or BankIt.<sup>10</sup> We briefly consider some of the requirements of such tools.

- *They must support reuse through controlled vocabularies:* A controlled vocabulary is a library of standardized, reusable definitions (‘concepts’), which can be searched by several criteria, for example, keywords and synonyms, type of parameter, and so on. A data submission tool must

have integrated access to up-to-date contents of multiple existing controlled vocabularies—most of which have evolved for distinct purposes. LOINC, mentioned above, is an example of a standard vocabulary for laboratory tests and clinical observations. If a user can (a) specify that a parameter is a clinical test or observation and (b) locate the parameter in LOINC by specifying one or more keywords, then many details do not need to be manually supplied, because pointing to a LOINC entry provides these details. An additional advantage of controlled vocabularies is that they use internal unique identifiers that are useful for cataloguing. These identifiers are mostly unambiguous, and their presence in data or metadata facilitates the operation of analytical programs. The use of vocabularies such as the Internal Classification of Diseases (ICD),<sup>11</sup> for example, is well known.

- *Reuse across multiple submissions by the same investigator:* Individual investigators submitting data to a phenotypic repository have specific research interests, and tend to work with the same kind of parameters over time. In cases where an investigator has submitted previous data descriptions, the software tool must index these descriptions by investigator, and permit retrieval of descriptions so that individual descriptors can be reused for a new submission.
- *Reuse within the same data set:* Several columns in a given data set may represent repeated measures of the same parameter over time: for example, one may record fasting blood glucose at baseline and at 3 months after baseline. The submitter must be able to point to a previously supplied description and reuse it for an additional column, changing

only the details about when the parameter was measured.

The software-engineering task of creating a user-friendly data submission tool is relatively straightforward. A much bigger challenge is arriving at a consensus among researchers as to what descriptors constitute a 'minimum acceptable' set with respect to different types of parameters. In the microarray world, for example, there is a definition called MIAME (minimum information about a microarray experiment).<sup>12</sup> A community effort towards a similar goal for phenotypic metadata is urgently needed. Individual research consortia, such as the Pharmacogenetics Network<sup>13</sup> or the Environmental Genome Project,<sup>14</sup> are in a good position to explore informatics approaches that will work within a consortium, but the time is probably ripe for the formation of *ad hoc* groups at various genomics meetings. Bioinformatics groups have an excellent track record of collaborative creation of open-source software, as in the various bio-software initiatives, and it is hoped that this article acts as the catalyst for such collaboration.

#### ACKNOWLEDGEMENTS

I thank Steve Sherry of the National Center for Biotechnology Information and Lisa Brooks of the National Human Genome Research Institute for discussions on the phenotype problem. This work was supported by NIH Grants U01 ES10867, R01 LM06843-01 and U01 CA78266.

#### DUALITY OF INTEREST

None declared.

#### Correspondence should be sent to

Dr PM Nadkarni, Center for Medical Informatics,  
Yale University School of Medicine, PO BOX  
208009, New Haven, CT 06520-8009, USA  
Fax: +1 203 764 6717  
E-mail: Prakash.Nadkarni@yale.edu

- 1 Berson A, Smith SJ. *Data Warehousing, Data Mining, and OLAP*. McGraw-Hill: New York, 1998.
- 2 Sherry S, Ward M, Kholodov M, Baker J, Phan L, Smigielski E *et al*. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001; **29**: 308–311.
- 3 National Center for Biotechnology Information. Haplotype Specifications. 2002. <http://www.ncbi.nlm.nih.gov/SNP/HapMap/index.html>. Last accessed on: 7/8/02.
- 4 Blamire J. Science at a Distance. <http://www.brooklyn.cuny.edu/bc/ahp/BiolInfo/GP/Definition.html>. Last accessed on: 7/7/02.
- 5 Fajans S, Bell G, Polonsky K. Molecular mechanism and clinical pathophysiology of maturity-onset diabetes of the young. *N Engl J Med* 2001; **345**: 971–980.
- 6 Microarray Gene Expression Data (MGED) Group. MGED Home Page. 2002. <http://www.mged.org/index.html>. Last accessed on: 7/8/02.
- 7 Regenstrief Institute. LOINC home page. 2002. <http://www.regenstrief.org/loinc/>. Last accessed on: 7/8/02.
- 8 Bakken S, Cimino J, Haskell R, Kukafka R, Matsumoto C, Chan G *et al*. Evaluation of the clinical LOINC (Logical Observation Identifiers, Names, and Codes) semantic structure as a terminology model for standardized assessment measures. *J Am Med Informatics Assoc* 2000; **7**: 529–538.
- 9 Briehl RW. Sickle cell hemoglobin. In: Dulbecco R, ed. *Encyclopedia of Human Biology*, 2nd edn. Elsevier Science (Academic Press): Amsterdam (New York), 1997. pp. 1–20.
- 10 Benson D, Karsch-Mizrachi I, Lipman D, Ostell J, Rapp B, Wheeler D. *Nucleic Acids Res* 2002; **30**: 17–20.
- 11 World Health Organization. *International Classification of Diseases*, 10th Ed: WHO: Geneva, Switzerland; 1998.
- 12 Microarray Gene Expression Data (MGED) Group. Minimum Information about a Microarray Experiment. 2002. <http://www.mged.org/Workgroups/MIAME/miame.html>. Last accessed on: 7/8/02.
- 13 Davis A, Long R. Pharmacogenetics Research Network & Knowledge Base: First Annual Scientific Meeting. *Pharmacogenomics* 2001; **2**: 285–289.
- 14 National Institute of Environmental Health Sciences. Environmental Genome Project. 2002. <http://www.niehs.nih.gov/envgenom/home.htm>. Last accessed on: 8/20/2002.