Recommendations on performance assessment scoring in ERNDIM qualitative proficiency testing schemes

Jim Bonham, Neil Dalton, Marinus Duran, Brian Fowler, Claus-D.Langhans, Rodney Pollitt, Christine Vianey-Saban and Viktor Kožich

Version agreed at the SAB 3.12.09

A. AIM

The purpose of this recommendation is to harmonize the evaluation of diagnostic proficiency of laboratories participating in qualitative ERNDIM Schemes

---

B. PRESENT STATUS OF QUALITATIVE SCHEMES

1. Schemes: ERNDIM presently runs five diagnostic proficiency testing schemes (Amsterdam, Basel, Lyon, Prague and Sheffield), two qualitative organic acid in urine schemes (Heidelberg and Sheffield) and one qualitative blood spot acylcarnitine scheme (London). Presently the Schemes are administered by individual centres which are responsible for sample collection and assessment of sample suitability, sample selection for each control cycle and distribution of samples, for collection of reports by individual laboratories, evaluation of these individual results and in the case of diagnostic proficiency testing also for ensuring discussion and final approval of scoring by scheme participants at an annual meeting.

The main purpose of all qualitative schemes is to evaluate the ability of the laboratory to establish or exclude a specific diagnosis of an inherited metabolic disease.
In **diagnostic proficiency schemes** the analytical approaches which have to be used are not specified. The laboratory has to select appropriate methods, obtain correct analytical results, evaluate them with respect to reference ranges or evaluate qualitative patterns, propose a likely diagnosis and suggest additional tests necessary for confirmation of the diagnosis. The spectrum of methods that should be available varies between different DPT centres. Each include creatinine, qualitative urinary dip tests, analysis of amino acids, organic acids and mucopolysaccharides while purines / pyrimidines are not obligatory in the Sheffield and Lyon DPT schemes and oligosaccharides are not obligatory in the Sheffield scheme.
If a laboratory does not run a method requested for participation, it is allowed to seek an analysis in a subcontracted lab (cluster lab) and the results for the entire sample are considered as being produced in the subscribed laboratory. However, a growing number of laboratories are becoming accredited and results from participation in DPT scheme may be used by accreditation bodies for assessing the laboratory´s performance. Therefore the policy of scoring results obtained by a subcontracted laboratory needs to be clarified.

The remaining qualitative schemes are method-oriented and the laboratories are expected to obtain correct analytical results, to recognize the characteristic diagnostic patterns, to make a diagnostic conclusion and to suggest additional test(s) necessary to confirm the diagnosis. Results obtained from a cluster lab are occasionally submitted and so far they have been scored as if they were produced by the subscribed laboratory, however this is felt to be inappropriate.

2. Samples in Schemes: So far only authentic urine or blood samples obtained from patients with a specific inherited metabolic disease (IMD) or from individuals with no known IMD are distributed in the qualitative Schemes (samples produced artificially by spiking urine or plasma with typical metabolites have not yet been used). The samples for distribution are

contributed by the organizers of each scheme as well as laboratories participating in the respective scheme and are stored frozen. Urinary samples are heat-treated in the presence of thiomersal, small aliquots are left at room temperature for 3 days to mimic transportation conditions and re-analyzed to ensure that typical analytes have not been altered by heat-treatment and storage at ambient temperature. Blood samples for distribution within the acylcarnitine scheme are collected locally or contributed by other participating laboratories. Samples are either stored immediately as whole blood at -80^{o}C and subsequently pipetted onto Whatman 903 filter paper prior to distribution or 40µl blood samples are pipetted onto Whatman 903 filter paper, dried overnight at room temperature, and then stored in plastic bags at -80^{o}C. The acylcarnitine profile is measured using an underivatised method on the day before despatch and again, following storage at room temperature, on the closing date for the circulation.

Samples are distributed once or twice per year using shipment at either room temperature or on dry ice by either postal service or door to door couriers.

3. Clinical information on samples: Clinical information on patients from whom the samples have been obtained varies between one short sentence and several sentences. The DPT organizer aims to provide the short clinical information that was provided by the referring physician at the time of the first referral when the diagnosis was established (i.e. the same information on which the laboratory was able to achieve the diagnosis). The minimum information contains age at presentation, sex, clinical symptoms and signs, age when sample was obtained and conditions when the sample was obtained (specific or non-specific therapy and sampling in acute crisis is mentioned **if felt relevant**).
For the acylcarnitine scheme very limited clinical information is provided based on the information provided by the referring physician/laboratory at the time of the first referral when the diagnosis was established.

4. Confirmation of diagnosis of samples circulated in Schemes: As laboratories are expected to reach a correct diagnostic conclusion the verification of the defect in the patient who provided samples is of utmost importance. The extent of diagnosis confirmation varies widely- from patients where the diagnosis was obtained at the metabolite level only compared with patients in whom the diagnosis was confirmed enzymatically or in patients where the diagnosis was confirmed by molecular genetic analysis. In practice, two levels of confirmation are the rule. Samples in which metabolite profiles are grossly abnormal and very typical for a single disorder or samples where the metabolite patterns can be observed secondary to non-genetic conditions are distributed without enzymatic/molecular genetic confirmation. Alternatively, samples exhibiting only minor abnormalities are usually circulated after confirmation has been obtained enzymatically or by DNA analysis.

There is at present no formal procedure (apart from clinical history) for excluding the possibility of an undiagnosed inherited metabolic disorder in subjects providing samples designated as “no known IMD”.

5. Evaluation of performance in DPT Schemes:
During previous years a harmonized scoring system has been adopted by all 5 DPT centres. For each sample 3 criteria are evaluated: analytical performance, interpretative proficiency and recommendations for further investigations. The scoring system is shown in the table below:

A	Analytical performance	Correct results of the appropriate tests	2
		Partially correct (or non-standard) methods	1
		Unsatisfactory or misleading	0
I	Interpretative proficiency	Good (diagnosis was established)	2
		Helpful but incomplete	1
		Misleading/wrong diagnosis	0
R	Recommendations for further diagnostic testing	Helpful	1
		Unsatisfactory or misleading	0

The **total score** per sample is calculated as a sum of these three criteria. The maximum that can be achieved is 5 points per sample. In the absence of any results the sample is scored as 0 points.

a) Scoring of analytical performance in DPT Schemes
Correct results (2 points) are awarded for clearly described and correct analytical results. This can be based on a semiquantitative evaluation of an abnormal level of metabolite(s) (i.e.not detected-very low/low-normal-elevated/grossly elevated) or a quantitative value evaluated correctly with respect to appropriate reference ranges. For qualitative analyses a description of the profile of analytes suggestive of a specific diagnosis or excluding it, is considered correct.

Partially correct analytical results (1 point) may be those obtained if not all of the necessary tests have been performed (e.g. quantitative mucopolysaccharide measurement without performing profile analysis of MPS), if the results are not sufficiently well described (e.g. “abnormal pattern” without suggesting the diagnosis in oligosaccharides or mucopolysaccharides profiling).

Unsatisfactory or misleading results (0 points) are given when the appropriate test has not been carried out or a key metabolite has not been detected or has been falsely identified.

b) Scoring of interpretative proficiency in DPT Schemes
The participants are expected to limit the number of possible diagnoses to the most likely ones.

Full score (2 points) is given for a correct diagnostic conclusion that is achievable by analyzing urine (i.e. the participants are not expected to “overdiagnose” if analyses with other fluids such as plasma or CSF are needed). If the analytical results are suggestive of a single possible diagnosis, this should be mentioned, preferably as a common name of the disease or as enzyme deficiency (e.g.argininosuccinic aciduria or ASL deficiency). Where several diagnoses are possible the differential diagnosis should be mentioned (e.g. methylmalonic aciduria due to mutase deficiency or cbl defects cbl A/cblB/cblD or secondary to B12 deficiency).

Helpful diagnostic conclusions (1 point) are those which will eventually lead to establishing the diagnosis. In the majority of cases these partially correct diagnostic conclusions lead to a group of diseases (e.g. diagnosis of mucopolysaccharidosis in the case of elevated MPS but without a clearly described profile or with assignment of a wrong type of MPS).

Wrong/misleading diagnosis (0 points) is assigned to participants for “overdiagnosis” in the case of no known IMD, completely inappropriate diagnosis or missed diagnosis in a patient with a known IMD.

Evaluation of interpretative proficiency is a challenging task for the organizers and the usefulness of the report to a non-specialist clinician is the guiding principle. Problems also arise when a lab did not perform appropriate tests (e.g. analysis of MPS) but still suggests a correct diagnosis of mucopolysaccharidosis based on the clinical description alone and recommends MPS analysis, i.e. should the interpretative proficiency be scored by 1 point even in the absence of any relevant analytical data? The scoring here is somewhat subjective but if the conclusions are considered “safe” and broadly correct then 1 point should be given.

c) Scoring of recommendations in DPT Schemes

An array of diagnostic procedures is usually proposed by participants and it is usually quite unusual to fail in recommendations. Scoring the recommendations brings many problems to the evaluators, e.g. redundancy of recommendations, detachment from practice (tests are recommended that will not be ordered in a real patient), non-specificity of recommendations, e.g. the very generic recommendation of enzymatic analysis or DNA testing without specifying an enzyme or gene, over invasive testing in cases of samples from patients with no IMD, quoting of an IMD database search.

d) Present consensus on satisfactory performance in DPT Schemes

At present six samples are circulated per year and the total score achievable is 30 points. In an initial effort to identify very poorly performing laboratories The ERDNIM SAB defined good / satisfactory performance initially as 15 points or more (until 2008) and raised this to 18 points or more in each yearly cycle to be applied form 2009 onwards. Laboratories failing to reach the agreed level are issued with an individualized warning letter which is aimed at helping them to improve the performance of the laboratory. At present approximately 5 to 10% of participants come into this category of poor performance although this might well increase as and when the threshold for good performance is raised.

6. Evaluation of performance in the qualitative schemes for urinary organic acids and blood acylcarnitines.

For each sample in the urinary qualitative organic acid schemes participants report under the headings of analytical findings, interpretation and recommended further tests. 2 points are awarded for a completely correct response identifying the key metabolites and making an appropriate interpretation, 1 point for a helpful but incomplete response eg a dicarboxylic aciduria and acylglycines with the conclusion of “a fat oxidation defect” in a urine from a patient with MCAD deficiency, 0 points for an unhelpful but not misleading response e.g. a dicarboxylic aciduria “requiring further investigation”. With nine samples per year a total maximum score of 18 points is possible. In contrast to the DPT schemes a penalty score of –2 points can be given for a potentially harmful misleading report such as the finding of "no abnormality detected" in the example already described.

The acylcarnitine scheme did not use any scoring system until 2008 and a certificate of participation was issued if results were submitted for at least one set of three samples from the two sets distributed each year.

C. HARMONIZED CRITERIA FOR RUNNING AND EVALUATING ERNDIM QUALITATIVE PROFICIENCY SCHEMES

The qualitative scheme organizers reached consensus on the following criteria and principles in striving for harmonisation of the schemes although some issues remain to be fully worked out.

1. Criteria for suitability of samples to be distributed

Suitable samples for qualitative schemes are non-modified urine and blood samples from human subjects with the following conditions;

- apparently healthy subjects under various conditions (including e.g. fasting, receiving special diets, foods);
- patients with diseases other than inherited metabolic disorders, e.g. with symptoms mimicking IMD and patients receiving treatment that may interfere in analytical procedures;
- patients with an inherited metabolic disorder that is detectable in urine or blood, according to the scheme. The urinary samples must contain typical analytes after they have been heat-treated at 56° C in the presence of 100mg / L thiomersal for one hour and storage at 25°C for 72 hours;
- dried blood spot samples containing typical concentrations of free carnitine and acylcarnitine species previously reported and/or observed to be diagnostically helpful.

Unsuitable samples are those from:

- patients with a known infectious disease that is transmissible by blood or urine;
- dilute urines with a creatinine concentration below 1 mmol/l,
- IMDs with metabolites that require non-standard preanalytical conditions (e.g. very unstable/volatile metabolites, metabolites that require special and non-standard solubilization procedures);
- IMD in remission with completely normal profile of relevant analytes.

If **in the future** obtaining sufficient authentic patients samples to run the schemes becomes impossible it may become necessary to produce samples by spiking normal urines. Although we now have great experience of the spectrum of metabolite patterns and how they vary in different disorders, and it is theoretically possible to produce plausible spiked samples, the unavailability or expense of compounds may make this difficult.

2. Clinical information on samples

Since appropriate clinical information may be very helpful in selecting analytical tests and in interpreting results a standardized format of clinical information is needed to ensure conformity both within and between schemes. It is assumed that the laboratory performing the diagnostic tests decides themselves which tests are to be performed. In the case of laboratories which only follow instructions to perform specific tests decided by referring physicians a list of tests that would usually be ordered by the referring physician for the particular sample and no clinical information could be provided. However this removes an essential element of testing the diagnostic approach of participating laboratories.
The clinical information should ideally mimic the situation when the sample arrived for the first time in a diagnostic laboratory and at which time the laboratory established the diagnosis using the given clinical details. The clinical information has to be composed of the following three parts:

a) sex and age of patient at time of the first referral to metabolic laboratory
b) clinical information at the time of the first referral, ideally the information given by the referring physician in writing and/or obtained by phone, listing the most important symptoms and signs of presentation but not those obtained after the diagnosis was made. Optimally information should cover about 2-3 lines of text and may be shortened if necessary.
c) data relevant to sample collection such as age of the patient when obtained, possible treatment which may modify the metabolite profile, and in the case of disorders with intermittent presentation information on the clinical state of the patient (e.g. well or ill).

3. Shipping of samples

To ensure consistency within and between Schemes the following recommendations have been adopted for sample shipping:

a) number of shipments per year- one or two

b) transportation by door to door courier service with shipment tracking.
c) Transportation at room temperature by the end of March to avoid excessively high temperatures (at least in the northern hemisphere)

4. Determination of target values and evaluating consensus values
For assessment of analytical performance and interpretative proficiency in qualitative schemes non-numerical target values (i.e. correct analytical finding and correct diagnosis, respectively) are used, for assessment of diagnostic recommendations non-numerical consensus values are used. Prudence of organizers is therefore needed in setting the target values although annual meetings of DPT Schemes allow for correction of both target and consensus values.

a) To obtain the target value for analytical performance the sample is analyzed by methods appropriate for the diagnosis in a reference laboratory, presently in the lab of the scheme organizer but if in future samples are stored and distributed by a central organization e.g. CSCQ, in laboratories determined as a reference lab by the SAB. They are evaluated with respect to the diagnosis and to previous analytical findings obtained by the referring laboratory if they exist. If no discrepancies are found between analytical results and the diagnosis a descriptive target is set by the organizers, e.g. elevated concentration of phenylalanine in urine, methylmalonic aciduria with homocystinuria, presence of SAICAR, elevated concentration of GAGs and increased proportion of heparansulfate, presence of valproate and its metabolites, physiological profile of urinary purines and pyrimidines etc. So far numeric target values have not been set due to the paucity of definitive methods, standards and calibrators in biochemical genetic testing.

b) The target value for assessing interpretive proficiency is i) the previously established diagnosis of a specific IMD or ii) the presence of typical non-specific findings such as those related to drugs or other diseases or conditions (e.g. fasting) or iii) the absence of these. For IMDs with unique and typical analytical patterns (e.g. argininosuccinic aciduria), confirmation of diagnosis at the metabolite level is acceptable. For IMDs with metabolite profiles suggestive of several disorders including non-IMDs, verification of diagnosis at the enzymatic or DNA level should be performed where relevant and possible before the sample is considered suitable for distribution. It is important to emphasize that the true enzymatic/DNA diagnosis may not be a target value achievable in qualitative schemes. For example in the case of MMA-CoA mutase deficiency the presence of highly increased methylmalonate concentrations in urine can occur in several other disorders such as the cblA/B/D defects, or as a secondary finding in vitamin B12 deficiency. The limitations of analyzing only one sample and not having access to repeated samples, samples of other body fluids and lack of relevant clinical information have to be considered when setting target values. In other words the target value has to be realistic and achievable in a majority of well performing labs.

c) Consensus for recommendations for further diagnostic tests is set by organizers after a critical review of submitted data. The recommendations should be ethically acceptable, efficient in distinguishing several diagnostic possibilities or crucial for confirming diagnosis in the patient. Consensus values for diagnostic recommendations are context-sensitive and may change from participant to participant for the same sample considering their results of analytical tests and conclusions. For example in a patient with mucopolysaccharidosis IIIA the correct recommendation of a laboratory that detected both elevated GAGs and an increased proportion of heparan sulfate should be to carry out enzymatic analyses relevant for MPS III. However the correct recommendation of a laboratory that detected only elevated GAGs but has not analyzed GAG fractions should be to perform electrophoresis or TLC fractionation.

5. Results obtained from a subcontracted laboratory

Certificates of performance issued by ERNDIM can be presented to accreditation bodies as a proof of successful participation in EQA Schemes. Therefore results obtained in a subcontracted lab cannot be considered as measures of laboratories' own performance. For qualitative organic acids and acylcarnitines schemes the laboratory must produce its own results and cannot send samples to a subcontracted laboratory. For DPT Schemes, the laboratory may send a sample to a subcontractor for a limited number of the tests deemed as essential for participation in the scheme. In this case this information has to be clearly stated and will be included in the certificate of participation. However the laboratory will receive full points for analytical performance any points for analytical performance, i.e. no points will be deducted and the total possible points will be 30 for all laboratories that participate. Also the interpretative proficiency and recommendations will be scored independently regardless of the place where analysis was performed.

6. Definition of Satisfactory performance in qualitative proficiency testing schemes

Diagnostic proficiency testing: the ultimate aim is to identify laboratories that clearly perform satisfactorily and to issue a certificate to them to that effect. As a first step in this direction until harmonized good performance assessment can be robustly defined it is appropriate to identify laboratories that clearly perform poorly. Starting on January 1, 2009 this is defined as those which fail to obtain 18 points or more in one year and they should receive a letter pointing this out and offering assistance for improving their performance. It has to be borne in mind that the degree of difficulty of samples may necessitate upwards or downwards adjustment of the absolute value.

Qualitative OA and acylcarnitines: these schemes are less complex and the method is prescribed by the scheme. The urinary organic acids scheme has long used a scoring system -2 to +2 awarded for each sample (see B-6 above). The possibility of applying minus points as a penalty emphasises the potential for harm as well as good. Individual samples vary greatly in difficulty so that, depending on the case-mix, the threshold for 'poor performance' may vary slightly from year to year. For 2007 both the Sheffield and Heidelberg schemes agreed on a minimum of 11 points out of a maximum of 18. This would be equivalent to 38 points from a maximum of 45 on the 5-point scale used in the diagnostic proficiency schemes. The same scoring system will be applied for the acylcarnitine scheme

7. Final comments and considerations for further harmonization.

It is clear that all of the ERNDIM proficiency schemes are harmonized by encompassing the three components of analytical performance, interpretive proficiency and recommendations for so that the actual numbering system employed is not critical. Both the DPT and qualitative schemes developed and fine tuned their numbering system over several years and these reflect both similarities and differences between the two types of schemes. Also acceptance by participants is high and it seems justified not to change the numbering systems used at the present time. One major discrepancy is the lack of penalty points for serious errors in the DPT schemes. Incidentally penalty points are also included in the EQA schemes operated by CEQA for cytogenetic and EMQN for molecular genetic testing.

Thus an important further step in harmonization which should be strongly considered is the inclusion of penalty scores for the DPT schemes.

D. IMPLEMENTATION

This recommendation will take effect from January 1, 2010 and the new system will be evaluated annually.