

Harmonisation of performance assessment in qualitative PT/EQA

Vivienne James



Definitions

- Harmony: the state of being in agreement or concord
- Harmonisation: the process of creating common standards
- EQA (Laboratory Medicine): the total process whereby the quality of laboratory results can be guaranteed
- PT: determines the performance of individual laboratories for specific tests or measurements and is used to monitor laboratories' continuing performance
- Qualitative measurement: Categorical measurement expressed by means of a natural language description
 - Nominal e.g. organism name/identity, genotype, presence/absence, positive/negative
 - Ordinal e.g. 1+, 2+, 3+ (can be ordered) but have no algebraic relationship
 - Presence/absence and positive/negative can also be considered ordinal , with just two values.



How can qualitative performance be assessed?

- Raw descriptive data can be categorised and comparison made between the categories
 - Organism identification to genus level, species level, species and serotype
- Comparisons can be interpreted
 - Number of results or % results in each category
 - Apply a numerical score to enable ranking



Benefits of a scoring system

- Simplify data – assist participants to assess their performance relative to other labs
- Enables comparisons between groups of laboratories
 - Method comparisons
 - Country comparisons
- Applying a numeric score provides a mechanism for monitoring performance over several rounds
- The score can be subjected to basic statistical analyses
 - Standard errors
 - Ranking

Harmonised performance assessment

Benefits

- easier to understand
 - by the participants and end users
 - customers, regulatory authorities and accreditation bodies
- comparable
 - Participants using different EQA providers so they can cover the range of testing undertaken by their laboratory
 - Potential for comparisons internationally

Problems

- Requires common scoring system
- Harder to understand
 - A complicated statistical approach may not be accessible to participants and their users generally dealing with qualitative results
- PT challenges not necessarily equivalent
 - Differences in the specimens sent
 - Need for defined standards for each property to be evaluated
 - Need for common sample specifications
- Too simplistic
 - Decision points
 - Interpretation

Scoring systems for qualitative schemes

Response to the EQALM questionnaire undertaken in 2013 to determine current practice in the field of laboratory medicine

- International drive to harmonise qualitative schemes
 - work item for the EURACHEM PT working group

1. Field (eg, clinical chemistry, immunology, microbiology)?
2. Parameter (eg, glucose in urine sticks, HIV antibodies)?
3. Are the results, nominal or ordinal ?
4. How the target value/correct result determined?
5. When scoring what tolerance is permitted for a full score (eg, target value only, +/- 1 interval) ?
6. Is performance evaluated over several surveys (eg, score based on 5 surveys) ?

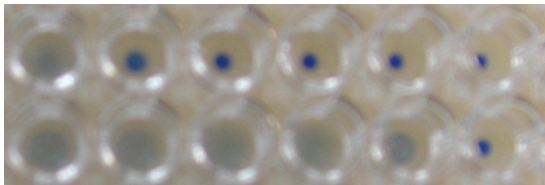
Questionnaire response

- Sent to all members >50
- Responses received from 35
 - 25 providers offer qualitative schemes
 - Providers represent 20 countries (18 European)
 - All major disciplines in Laboratory Medicine were represented
 - 18 providers score some or all their qualitative schemes

Scored qualitative schemes

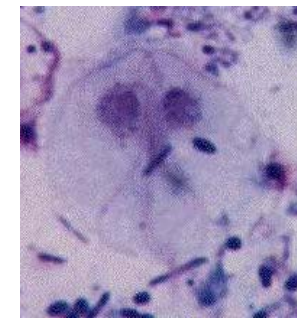
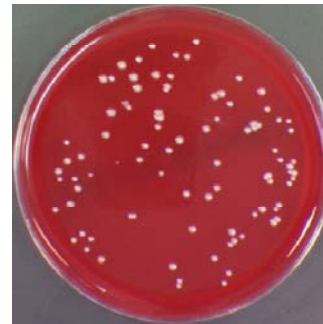
Ordinal results

- Target only
- Target +/- one interval
- Target value with weighted score



Nominal results

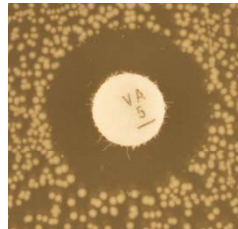
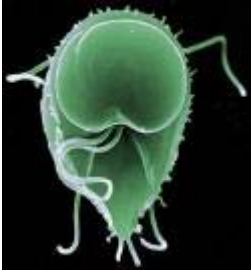
- Survey dependant
- Target value only
- Target value with weighted score



Scoring mechanism was provider dependant

Non scored schemes

- 11 providers offer qualitative schemes (covering all disciplines) that are not scored
 - 3 of these providers evaluated performance annually over several rounds
 - 1 required 75% correct results for successful participation
 - 9 providers did not score schemes with nominal results
 - 2 of these evaluated performance over several rounds
 - 3 providers did not score schemes with both nominal and ordinal results
 - 1 of these evaluated performance over several rounds
 - 1 provider did not score a scheme with ordinal or nominal results
 - 1 provider did not score schemes with ordinal results
- For the parameters offered that were not scored other providers did score



Summary

- Majority of qualitative tests have nominal results
 - parameters that are considered to have ordinal results by one provider are considered nominal by others even within the same discipline
 - relates to the clinical action and/or interpretation
- If the results are nominal the target value is required for a full score
 - however some providers weight the score depending on the clinical importance of the result
- If the results are ordinal or can be ordinal or nominal a tolerance of one interval is accepted by some providers

Unanswered questions from the survey

- Does whether scoring is used reflect the approach to quality in the country that the scheme is based (participants will actively review all EQA results) ?
- Would providers introduce a score if there was a recommendation on suitable scoring systems for qualitative schemes in ISO 13528?



Types of scoring schemes

- Single response (diagnosis)
 - Immune/non-immune; normal/abnormal
 - Weighted degrees of how right, partial identification
 - Top marks – the better you are the higher your score
 - 2 fully correct, 1 partially correct, 0 wrong, -1 grossly misleading
 - Penalty points – the better you are the lower your score
 - 0 fully correct; 1 partially correct; 2 incorrect; 3 grossly misleading
- Multiple response (differential diagnosis)
 - Likelihood of each diagnosis
 - Risk e.g Down's syndrome in foetus



UK NEQAS Haematology 1

Nominal scoring (nature of analytes)

Target:

- Reference result *or* expert panel *or* consensus of participants' results
- Consensus must be agreed by set minimum % of participants
- Each specimen scored individually
- Look up table for scoring results of each specimen: e.g. Correct = 0, Incorrect = 35/50 etc points depending on significance
- Score for a single survey = sum of scores for specimens in that survey
- Cumulative scoring – usually approx 6 months time window for surveys for which results have been returned (i.e. non-returns are skipped):
 - Score for each survey truncated to 50, and then summed over 3 surveys, *or*
 - Score per specimen summed over 6 consecutive samples
- Unsatisfactory performance = cumulative score of 50 in one time window
- Persistent unsatisfactory performance = cumulative score of 100 in one time window
- Unresolved persistent unsatisfactory performance = cumulative score of 150 in one time window

Barbara De la Salle 2013



UK NEQAS Haematology 2

- Flexible
 - Applicable to diverse analytes
 - Penalty graded according to clinical impact
- Easy to automate
 - Running score can be presented graphically
- Same action points across all schemes
- Same action points as quantitative schemes
- Truncation avoids unfair penalty where necessary



IEQAS: Histopathology Scoring System

- Parameter
 - Histopathology slide (12/distribution) for nominal diagnosis
- Target value and Tolerance
 - Peer review after data submission
 - Correct/acceptable diagnosis ↗ submitted by ↑ 80% members
 - Case/material unsuitable? (each response arbitrarily scored as 1)
 - Individual 'other' responses acceptable with reduced score?
- Scoring (each case)
 - 1 Correct or acceptable response
 - 0.5 Incomplete, deficient or minor error of no clinical significance
 - 0 Incorrect or no response
 - (up to 3 differential diagnosis with weighting - accepted but not encouraged)
- Performance evaluated over several surveys?
 - Member score ↓ 2.5% for 2 out of 3 successive circulations (PSP)

Example: Hypothetical Case No 3

Diagnosis	Weighting (only if necessary; must add up to 10)
1. Basal cell carcinoma	8
2. Trichoepithelioma	2
3.	

- at least 80% of responses are of trichoepithelioma and this response is judged correct and is scored 1
- Benign skin tumour is considered only partly correct and is scored 0.5
- Basal cell carcinoma is considered incorrect and is scored 0

Scoring of response no. 3 = 0.2

First differential diagnosis, basal cell carcinoma = 0, weighting 8 out of 10 = score $0 \times 0.8 = 0$

Second differential diagnosis, trichoepithelioma = 1, weighting 2 out of 10 = score $1 \times 0.2 = 0.2$



Example of harmonisation of evaluation ERNDIM schemes

European Research Network for evaluation and improvement of screening, Diagnosis and treatment of Inborn errors of Metabolism

- Harmonisation of 5 diagnostic PT schemes to evaluate the ability of the testing laboratories to establish or exclude a specific diagnosis of an inherited metabolic disease
- Analytical approach is not specified by the PT provider
- Testing labs have to select appropriate methods, obtain correct analytical results, propose a likely diagnosis or suggest additional testing to confirm the diagnosis

○ Bonhan et al 2009



Harmonised scoring system

- 3 criteria – maximum score 5

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



Harmonised criteria for running and evaluating the schemes

- agreed by the scheme organisers

- Sample suitability
 - Non modified clinical samples
- Appropriate clinical information
 - Age/sex/clinical info at time of first referral/treatments
- Determination of target values
 - Must be realistic and achievable
- Consensus recommendations for further tests

Definition of satisfactory performance

First identify poor performance

Actual numbering (scores) not critical

Recognise there are differences between schemes

Participant familiarity with scoring schemes

Lack of justification for change




Alternatives to harmonisation

- Clarity by each provider that they have a scientifically defensible process for assigning the result for each challenge
- Clarity about the acceptance limits (if any) for acceptable performance for a specimen, round and/or performance period



View point of one EQA provider

- 'One word of caution - our scoring system is complex and sophisticated and has been developed in line with that for our quantitative schemes. Although international harmonisation is a 'good thing', I am concerned at the possible dumbing down of well developed schemes to fit with those providing a less elaborate service.'



Are statistics really required for qualitative schemes?

- Without statistics: assessment (a) needs criteria (c) on which to base success or failure
 - e.g. if a is % correct and c is 80%
 - $a > 80\%$ ok but $a < 80\%$ not ok
- With statistics: statistics needs a formula (f) to enable calculation
 - assessment (a) needs criteria (c) on which to base success or failure
 - statistics c relies on numbers(n) derived from f
 - $a \equiv f(n_c) > c$ ok but $f(n_c) < c$ not ok



Final considerations (1)

- Is harmonisation across all sectors for all schemes:
 - Desirable?
 - Achievable?
- Schemes need to be comparable before harmonised performance analysis can truly be used to compare participant performance
- If a harmonised scoring system was recommended a starting point may be to apply this to new schemes
- In the field of laboratory medicine a harmonised approach as described for the ERDIM schemes might be achievable for small schemes where real clinical samples with associated medical histories/diagnoses are available
- True comparability could be achieved through an internationally available EQA from a single provider!



Final considerations (2)

- Scoring is a tool to allow comparison of your results with that of the 'average laboratory'. It helps to bring individual discrepancies to your attention
- For qualitative PT/EQA where the result is either right or wrong the discrepancy is obvious and the reasons for the failure should be investigated.



Acknowledgements

- EQALM members
 - Barbara De la Salle
 - Hazel Graham