# Measurement of uncertainty in veterinary diagnostic testing

**Axel Colling, March 2010**

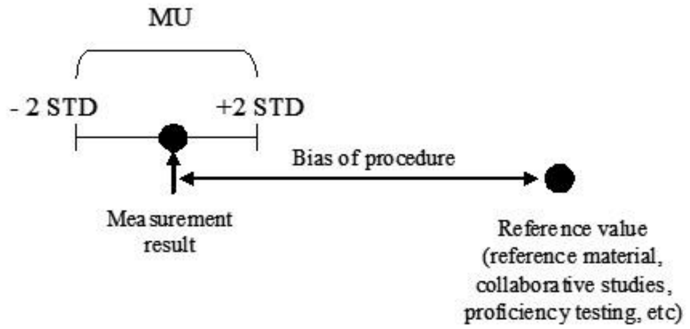
MU provides quantitative estimates of the level of confidence that a laboratory has in the analytical precision of test results, and is therefore an essential component of a quality system for veterinary diagnostic testing laboratories.

**Background**

The performance of a diagnostic tests is defined by two independent measures: precision and accuracy. Precision refers to the closeness of agreement among repeated measurements of the same sample under prescribed conditions and accuracy to the closeness of agreement between the result of measurement and the value of the analyte measured (Dybkaer et al. 1995). Every measurement made has an error associated with it and without a quantitative statement of the error it lacks worth. The parameter that quantifies the boundaries of the error of the measurement is called measurement uncertainty or MU (NATA, 2002). The OIE quality standard, 2008 defines *Measurement uncertainty as a parameter associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the analyte measured*.

In the context of diagnostic testing, MU provides quantitative estimates of the level of confidence that a laboratory has in the analytical precision of test results, and is therefore an essential component of a quality system for veterinary diagnostic laboratories. MU can be regarded as a combined measure of precision and bias, where precision measures the ability to repeat the result each time the same sample is tested and bias measures the ability to produce a 'true' result, as shown in Figure 1 (NPAAG, 2007).

**Figure 1: Measurement uncertainty (MU) as a combined measure of precision and bias**


MU
- 2 STD
+2 STD
Bias of procedure
Measurement result
Reference value (reference material, collaborative studies, proficiency testing, etc)

Toussaint et al (2007) offer the definition: *The uncertainty of measurement of a test results is the probability of not observing the same qualitative test result when retesting of the same sample*.

Validated methods provide information about precision for example repeatability and reproducibility and accuracy, and analytical and diagnostic sensitivity and specificity within established limits. Therefore MU is an important aspect of test validation but can not replace it.

Currently, MU is used for test methods that produce quantitative results, e.g. optical densities (OD), percentage of positivity or inhibition (PP, PI), titers, cycle threshold (CT) values. This includes tests, where numeric results are calculated and then are expressed as a positive or negative result at a cut-off value. Suitable statistical measures to express MU are mean values plus/minus 2 standard deviations (95% CI), relative standard deviation (rsd) or coefficient of variation (CV).

MU is applied to the analytical procedure and not to pre- or post-analytical errors such as sample suitability, collection, transport and transcription or reporting errors. Also excluded are interfering biological factors of the animal such as sex, breed, co-infection with other agents, age, body condition, pregnancy and immunity. Most of these components are included in the validation process.

There are two main approaches to estimate MU:

1) The ‘components’ or ‘bottom-up’ approach identifies all sources of uncertainty individually in a ‘fish-bone’ diagram. Chemical and physical testing laboratories tend to follow this approach because potential sources of uncertainty are usually readily identifiable, and their magnitudes can be estimated and combined. There are also published attempts to validate this approach in the medical testing field. For example, for serology, the uncertainties for time, temperature, volume, reading (OD), operator and reagent batch were identified to estimate the overall MU of the method (Dimech et al. 2007). The advantage of this approach is that the major sources of uncertainty are clearly identified and weighted individually. The results from Dimech et al. 2007 indicated that reagent batch-to-batch, lab-to-lab and operator variation contributed significantly to the total variation whereas reading, volume and temperature contributed to a lesser extent. The disadvantage is that it is a time-consuming process because it requires a complex statistical model and repeated measurements of each component.

2) The ‘control sample’ or ‘top-down’ approach is suitable for medical and veterinary diagnostic test methods because of the availability of quality control samples, which can be used to monitor whole-of-procedure-performance and directly estimate the combined MU of the test procedure. Upper and lower limits to approve or reject MU will depend on the purpose of the test. If the MU goal is not met it may be necessary to analyse the procedure to identify and modify uncertainty sources using the bottom-up approach. The advantage of this approach is the availability of repeatability data in diagnostic testing laboratories and simple calculations. The disadvantage is that the result is a global MU for the entire procedure and it fails to differentiate between individual contributing components.

Alternatively, the method characteristics approach, where performance data from a valid collaborative study are used as combined uncertainties (other sources may need to be added). Laboratories must meet defined criteria for bias and repeatability for the MU estimates to be valid. These should be larger than would be obtained by competent laboratories using their own control samples or components model (Dimech et al 2006).

The authoritative reference for MU is the Guide to the Expression of Uncertainty in Measurement (GUM), 1995. The GUM was specifically developed for calibration and testing laboratories in the field of analytical chemistry and physical testing (e.g. mechanical, electrical, temperature) and does not address the special nature of much of quantitative medical and veterinary testing. The purpose of this document is to provide guidance for the application of MU in veterinary diagnostic testing.

# MU in veterinary diagnostic testing

Because the measurement process is not entirely reproducible, there is no exact value that can be associated with the measured analyte. So the result is most accurately expressed as an estimate together with an associated level of imprecision. This imprecision is the MU. MU is limited to the measurement process. It is not a question of whether the measurement is appropriate and fit for whatever use to which it may be applied. It is not an alternative to test validation, but is rightly considered a component of that process.

Because there is MU associated with serological and other diagnostic measurements, the question is how to best estimate the MU? The framework against which MU must be applied is given by the standard against which the laboratory is accredited. The ISO/IEC 17025:2005 standard, *General requirements for the competence of testing and calibration laboratories* specifies the standards to which laboratories must operate if they are to achieve accreditation. For MU, there are the following requirements:

- MU as a part of validation (5.4.5.3)
- A decision (5.4.6.2)
- A procedure (5.4.6.2)
- Estimation (5.4.6.2 )
- The right impression in reporting (5.4.6.2)
- The use of experience and validation data (5.4.6.2)
- Consider what the test does and what the client needs (5.4.6.2 Note 1)
- MU for important components (5.4.6.3)
- Report MU relative to specification limit (5.10.3.1).

While there is considerable guidance from the Standard and international accreditation authorities as to what is required, only recently particular approaches for ELISAs (SCAHLS, 2007, Toussaint et al., 2007) and molecular tests (Goris et al., 2009) have been described for calculation of MU. The approach to applying MU principles to measurements in veterinary laboratories is still evolving, and that this document only provides an overview of the discipline-specific aspects with some examples.

A 'top down' approach uses precision and accuracy measurements to represent the MU. This approach recognises that the components of precision will be manifest in the ultimate measurement. So monitoring the precision of the measurement over time will effectively show the combined effects of the imprecision associated with component steps. The advantage of this approach for serology is that it avoids the need for empirical estimation of the effects of each of the many steps in for example, an ELISA. The disadvantage is that there must then be some other estimation of the effects of the imprecision in the component steps. Fortunately, the standard allows some latitude here. Part 5.4.6.2 of the standard allows consideration of 'the nature of the test method' and reasonable estimates 'based on knowledge of performance of the method and on the measurement scope'. So it should be possible to evaluate the steps of an ELISA and, in part, give quantitative assessments (e.g. pipette accuracy, plate reader precision) and apply qualitative assessments based on experience for other considerations (e.g. effect of temperature and pH within the stated range).

Australia's National Association of Testing Authorities, NATA Technical Circular Number 2 *Uncertainty of measurement in biological, forensic, medical and veterinary testing* suggests:

> The inter-batch precision of a method will often provide the best estimate of the uncertainty of measurement.

This approach is supported by the 2008 OIE Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, which provides more specific information as to how MU may be estimated using a top down approach that might be particularly applicable for serology:

> An example of the use of standard deviation to express uncertainty is the allowed limits on the test run controls for an enzyme-linked immunosorbent assay, commonly expressed as +/– n SD.

and

> A traditional control sample procedure, run many times by all analysts and over all shifts, usually covers all the major sources of uncertainty except perhaps sample preparation. The variation of the control samples can be used as an estimate of those combined sources of uncertainty.

For most antibody detection tests, it is important to remember that the majority of tests are measurements of antibody activity read relative to a threshold against which a dichotomous interpretation of positive or negative is applied. This is important because it helps to decide where application of MU is appropriate. In serology, uncertainty is frequently most relevant at the threshold between positive and negative determinations. The relative difference between rising titres may be used as a decision point in the inference of recent infection, so the uncertainty associated with defining a significant difference in titres should also be estimated where appropriate.

# Example

A limited data set from a competitive ELISA for antibody to avian influenza virus is used to illustrate a simple example of a "top-down" approach for serology. A low positive control sample was used to calculate MU at the cut-off level.

In serology, participation in proficiency test programs may provide at least a consensus mean against which accuracy (observed/expected) may be established by evaluating the difference of results between laboratories. Both the observed and expected values are estimates with associated uncertainties. If the measurement is relative to an internal standard only, there is no recognized standard against which to gauge accuracy. In such cases, bias is self-compensating and does not need to be considered. There is currently no available proficiency test program for the assay referred to in this example.

As the uncertainty is to be estimated at the threshold, which is not necessarily the reaction level of the low positive control serum, the relative standard deviation, rsd (or coefficient of variation, if expressed as a percentage), provides a convenient transformation:

$\text{rsd}(x) = \text{sd}(x)/(x)$

To simplify assessment, the transformed result is regarded as the assay output result (x). In the case of this example, a competitive ELISA, results are standardised by forming a ratio of all optical density (OD) values to the OD result of a non-reactive (negative) control ($OD_N$). This ratio is subtracted from 1 to set the level of antibody activity on a positive correlation scale, the greater the level, the greater the calculated value. This adjusted value is expressed as a percent now referred to as the percentage inhibition or PI value. So for the low positive control serum ($OD_L$), the transformation is:

$PI_L = 100 \text{ X } [1-\{OD_L/ OD_N\}]$

The relative standard deviation becomes:

$\text{rsd}(PI_L) = \text{sd}(PI_L)/PI_L$

A limited data set is shown below. Ideally in the application of this “top down” method, a large data set would be used, which would enable accounting for effects on precision resulting from changes in operator and assay components (other than the control serum).

| **Test** | **PI (%)** |
|---|---|
| 1 | 56 |
| 2 | 56 |
| 3 | 61 |
| 4 | 64 |
| 5 | 51 |
| 6 | 49 |
| 7 | 59 |
| 8 | 70 |
| 9 | 55 |
| 10 | 42 |
| | |
| Mean | 56.3 |
| Std Dev (sd) | 7.9 |
| Assays (n) | 10 |
| rsd = sd/mean | 0.14 |

From the limited data set:

rsd ($PI_L$) = 7.9/56.3 = 0.14 (or as coefficient of variation = 14%)

If uncertainty about accuracy is to be included, it is added at this stage by combining the squared rsd values, i.e.:

Combined uncertainty $u(x)/(x) = \sqrt{[\Sigma( \text{rsd } (x)^2]}$

However, for many serology assays, this information is not readily available and will not be considered in this example.

Expanding the uncertainty by multiplying the rsd ($PI_L$) by a factor of 2, allows the calculation of approximate 95% confidence levels around the threshold value, assuming normally distributed data

Expanded uncertainty $U_{(95\%CI)}$ = 2 X rsd = 0.28

This estimate can then be applied at the threshold level (in this case at PI = 50%)

95% CI = 50 ± (50 X 0.28)

= 50 ± 14%

Interpretation: any positive result (PI > 50%) that is less than 64% is not positive with 95% confidence. Similarly, a negative result (PI < 50%) that is not less than 36% is not negative at the 95% confidence level.

This zone of lower confidence for interpretation may correlate with the “grey zone” or “inconclusive/suspect zone” for interpretation that should be established for all tests.

The standard does not require reporting of MU unless:

- it is relevant to the validity, application or interpretation of the test results
- a client’s instruction requires reporting
- the uncertainty affects compliance to a specification limit.

In practical terms, this will require reporting of equivocal results, retesting of samples or otherwise advising clients of the proximity to the threshold.

Again, it is important to reaffirm that the MU does not represent test validity. For example, a threshold may be set high or low depending on the diagnostic characteristics required for the assay. The MU will apply in either case indicating reduced confidence for discrimination about that arbitrary threshold as test results fall within that zone. Depending on how the threshold has been set and other aspects of test validation, this uncertainty may have little or no impact on the interpretation of the result.

The top-down approach should be broadly applicable for a range of diagnostic test including molecular tests. For the calculation of tests using a typical two-fold dilution series for the positive control (such as in a VNT, CFT, HI) should be carried out on the geometric mean titre (e.g. mean and sd of log base 2 titre values) of the positive control serum. Relative standard deviations based on these log scale values may then be applied at the threshold (log) titre, and finally transformed to represent the uncertainty at the threshold. However, in all cases, the approach assumes that the variance about the positive control used to estimate the rsd is proportionally similar at the point of application of the MU, for example at the threshold. If the rsd varies significantly over the measurement scale, the positive control serum used to estimate the MU at the threshold should be selected for an activity level close to that threshold.

See also, worked examples of MU

For quantitative real-time PCRs (qPCR) replicates of positive controls with their respective cycle threshold (CT) values can be used to estimate MU using the top-down approach. Information about the most suitable extraction procedure and matrices, swab, blood etc. is normally obtained during the validation process and should be available prior to any MU assessment. Goris et al. (2009) estimated the MU of two real-time RT PCR methods for FMD (rRT-PCR FMDV 3D and FMDV 5'UTR assay) in spiked blood samples by testing the same dilution series ($10^0$-$10^{-6}$) in duplicate on 10 different occasions, e.g. over 10 days by 3 different operators to include all likely sources of laboratory variation. Hence, 20 individual CT-values were obtained for each viral dilution. Probability values for any possible result falling within the detection range of the assays were calculated. Uncertainty about a positive test result was defined as the probability of not observing the same qualitative test results, e.g. positive in this case, when retesting the sample a second time. The interpretation of results obtained for both assays was that any sample with a CT-values below 37.7 and 37.8 in the rRT-PCR and FMDV 5'UTR assay, respectively, had a probability of 99.0% of scoring positive upon retest (certainty). Similarly, any sample with a CT value of 41.4 in the FMDV 3D assay and 41.6 in the 5'UTR assay had a probability of retesting negative of 10% (uncertainty).

Results were also used to assess intra- and interassay variation. Interestingly, the intra-assay variation of the FMDV 3D assay was higher than the interassay variation. The FMDV 3D had a higher intra- and interassay variation (CV =4.6 and 3.5) than the FMDV 5' UTR assay (CV 1.9 and 2.7). The authors concluded that similar estimates should be done for other modern molecular and immunological techniques such as nucleic acid sequence based amplification (NASBA), loop-mediated isothermal amplification (LAMP) and ligation assays proximity (PLA).

## Concluding remarks

Quality oriented laboratories are always interested in monitoring the performance of their diagnostic tests for continual improvement. The use of internal quality controls over a range of expected results has become part of daily quality control and quality assurance operations of accredited facilities. Results provide relevant information about different aspects of repeatability, e.g. intra- and inter-assay variation[1], intra- and interoperator variation, intra- and

[1] If reference standards or calibrated controls against reference standards are used.

inter batch variation and inform about the level of robustness of a test procedure. The level of variation of a test result becomes increasingly important the closer the test values is to the cut-off value used to designate a test result as positive or negative. On the other hand, we normally have little doubt about test results that are on the extreme ends of the measurement scale and tend to call these results 'strong positive' or 'strong negative'. It is good laboratory practice to define a range of inconclusive, intermediate, suspicious, borderline, grey zone or equivocal test values falling between the positive and negative cutoffs. Greiner et al. 1995 describe an intermediate range and respective confidence intervals for test values for serodiagnostic tests falling around the cut-off. This range of values is considered as a borderline range for the clinical interpretation of test results. The proportion of the measurement range that gives unambiguous test results can be expressed using the intermediate range as the valid range proportion. In this context the relevant information that MU provides for diagnostic test results is that it gives an estimate about the extension of this range of values around the cut-off.

Once this range has been established, the diagnostician need to develop a test algorithm which describes how to follow-up samples which fall in the MU range. This can be a retest of the same sample or of a second sample and depends on the purpose of the test and its performance characteristics, in particular precision and accuracy[1]. Results from internal quality controls can be easily applied to estimate MU using a top-down approach with a minimum of additional testing and fulfil the requirements of ISO 17025.

# References

A2LA, 2009, *Policy on estimating measurement uncertainty for life science testing labs* (accessed 28 August 2009), American Association for Laboratory Accreditation.

Dybkaer, R., 1995, Result, error and uncertainty. *Scandinavian Journal of Clinical and Laboratory Investigation* 55: 97–118.

Dimech, W., Francis, B., Kox, J., Roberts, G., 2006, Calculating uncertainty of measurement for serology assays by use of precision and bias. *Clinical Chemistry* 52(3): 526–529.

Eurachem, 2000, *Guide to quantifying Uncertainty in Analytical Measurement*, Second Edition, Eurachem Secretariat, as Secretary, Nick Boley, LGC Limited, Middlesex, UK.

Goris N., Vandenbussche F., Herr, C., Villers, J., Van der Stede, Y. & De Clercq, K., 2009, Validation of two real-time PCR methods for foot-and-mouth disease diagnosis: RNA-extraction, matrix effects, uncertainty of measurement and precision, *Journal of Virological Methods* 160: 157-162

Greiner, M., Sohr, D. & Goebel, P., 1995, A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests, *Journal of immunological methods* 185: 123–132.

ISO/CASCO, 2005, General requirements for the competence of testing and calibration laboratories, AS ISO/IEC 17025-2005, Second Edition, ISO/CASCO Committee on conformity assessment, International Organization for Standardization.

JCGM, 1995, *Guide to the expression of uncertainty in measurement (GUM)*, ISO/IEC Guide 98:1995, Joint Committee for Guides on Metrology, International Organization for Standardization.

OIE, 2008, Infectious disease, in: *OIE quality standards and guidelines for veterinary laboratories—2nd Edition*, World Organisation for Animal Health.

OIE, 2009, Manual of diagnostic tests and vaccines for terrestrial animals, Chapter 1.1.3. Quality management in veterinary testing laboratories, d) Uncertainty, pp. 27–33, World Organisation for Animal Health (accessed 28 August 2009).

NATA, 2002, Assessment of uncertainties of measurement for calibration and testing laboratories, National Association of Testing Authorities, Australia (accessed 28 August 2009).

NPAAG, 2007, *Requirements for the estimation of measurement uncertainty*, National Pathology Accreditation Advisory Group, Canberra (accessed 28 August 2009).

SCAHLS, 2009, *Worked MU examples*, Sub-Committee on Animal Health Laboratory Standards, Canberra.

Toussaint, J.F., Assam, P., Caij, B., Dekeyser, F., Knapen, K., Imberechts, H., Goris, N., Molenberghs, G., Mintiens, K. & De Clercq, K., 2007, Uncertainty of measurement for competitive and indirect ELISAs. *Revue scientifique et technique (International Office of Epizootics)* 26(3): 649–656.