Review of a Proficiency Testing program for clinical markers

Hugo Gasca Aragón,
Centro Nacional de Metrología, Querétaro, México.
hgasca@cenam.mx

Introduction

Comparing against peers has been a common practice in proficiency testing (PT) for clinical field laboratories in Mexico, specifically by using the participants' consensus as the assigned value. However this practice introduces unbounded risks and consequences for the final users of these results. In order to show the benefits of using an assigned value with a higher metrological level a PT program for clinical laboratories using certified reference materials (CRM) was conducted by CENAM between 2002 and 2006 [1,2,3]. The results of this program are reviewed in this presentation.

Method

The program's original scope was to analyze five clinical key markers (calcium, glucose, cholesterol, creatinine, uric acid) in human serum. All participants received one sample and were asked to obtain 10 independent measurements during 2002 and 4 independent measurements thereafter. Participants were not asked to estimate measurement uncertainty, uncertainty due to repeatability conditions may be underestimated by the sample standard deviation. Only three measurands were selected for this study, the measurands' certified concentration value varied in the following ranges; glucose: 82.9 - 90.82 mg/dl, cholesterol: 156.3 – 161.18 mg/dl, creatinine: 0.74 – 0.75 mg/dl. All reference material were produced and certified by CENAM. CENAM used Isotopic Dilution method considered a potential primary method. The relative standard uncertainty [4] of the certified values ranges from 0.5% to 4%.

The participants' consensus was obtained by using non-robust (mean and standard deviation) and robust statistics [5] (median and scaled mad). The robust estimate avoids any criticism about declaring some participants as outliers. The assigned value obtained by the participants' consensus was compared against the certified reference value and the zeta (ζ) score was obtained for each one in both cases. The results based on robust statistics are shown in Figure 1, the results based on non-robust statistics are shown in Figure 2.

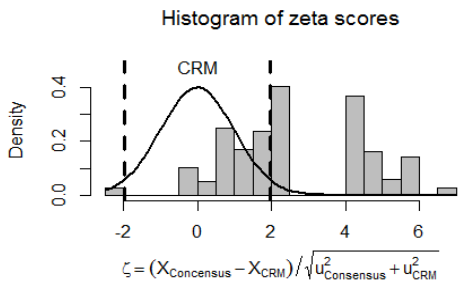

Figure 1. Assessing the participants' consensus estimated with robust statistics against the certified value.

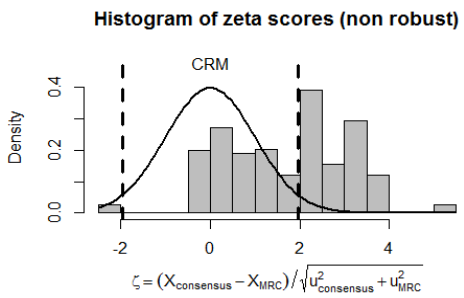

Figure 2. Assessing the participants' consensus estimated with non-robust statistics against the certified value.

This result shows the presence of bias for the specific sample of participants at least for some of the measurands. Note that the non robust estimates appear less dispersed or less biased than the robust estimates. This sheds some light about the risk that robust techniques may be led by outlier data.

In order to make inferences on the laboratories' whole population's measurement capability a simulation study was conducted. A non-parametric bootstrap [6] technique was used by resampling the PT results with replacement. This non-parametric resampling technique makes no assumptions about the underlying distribution of the random errors, hence it plays no role on the distributional properties of the results.

Simulation data was fitted with a one way random effects model in order to estimate the between PT exercises variance for each simulated participant. This component of uncertainty $\sigma_{mt\%}$ accounts for intermediate measurement precision [4,8,10] and is included in the uncertainty budget for each participant.

The distributions of the simulation's results are tested under two conditions: (a) using all the simulated data and (b) using the simulated data filtered for large variability (>20%) and bias.

Results

Modified target plot [11] with the simulation results were prepared for each measurand. Target plots were modified in two aspects:

i. The plotted quantities are relative to the reference value, in this order we use the percentage differences [7],
$$D_{\%} = 100(x - X_{CRM})/X_{CRM}$$

ii. The acceptance region is an arc-shape. This arc-shaped region is approximately the region with higher probability for the true joint distribution of both parameters under normality and independence assumptions [12]. The diagonal lines satisfy the equation
$$|D_{\%}| = \Phi^{-1}(0.975) \cdot \sigma_{mt\%}.$$

iii. The uncertainty of the percentage difference was estimated as
$$\hat{u}[D_{\%}] = \sqrt{\hat{\sigma}^2_{mt\%} + \left(100 \cdot \frac{u_{CRM}}{X_{CRM}}\right)^2}.$$

Figure 3 shows the modified target plots for glucose and creatinine.

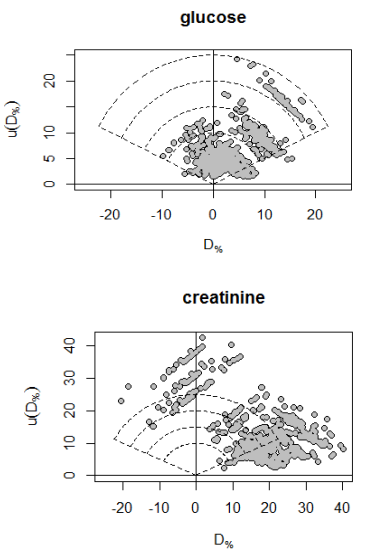

Figure 3. Modified target plots of the simulation results for (top) glucose, (bottom) creatinine. Each point is the mean and uncertainty of the participating laboratory over time.

Discussion

For glucose about 8% of the population of laboratories appears to have biased results, for cholesterol about 2% of the population of laboratories and for creatinine about 44%. Table 1 summarizes the distribution of the uncertainty of the unbiased results for each measurand.

Uncertainty of the percentage difference	% results Glucose	% results Cholesterol	% results Creatinine
$20 < u[D_{\%}]$	8	3	38
$15 < u[D_{\%}] \leq 20$	1	1	13
$10 < u[D_{\%}] \leq 15$	19	22	5
$u[D_{\%}] \leq 10$	64	72	<1

Table 1. Summary of the distribution of estimated uncertainty excluding simulation results suspected as biased. The use of CRM warrants the contribution to these uncertainties is less than 4%.

Mexican national policy [13] states a maximum of 5% for relative bias and relative uncertainty. The results for creatinine shown in Table 1 suggests a clear opportunity for improvement for the whole population of clinical laboratories and IVD kit producers.

Figure 4 shows a normal quantile-quantile plot of the zeta (ζ) scores for glucose of the simulation study under condition (b). Remarkably there is no evidence of non-normality although this distributional property had not been induced by the applied non-parametric resampling techniques.

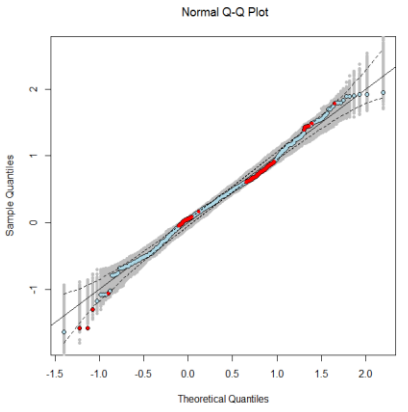

Figure 4. Normal qq-plot of the glucose simulation study under condition (b) (representing about 16% of the trials for glucose). The dots are the simulated values, the light blue dots are within the 95% confidence band for a normal distribution and the red dots are outside the 95% confidence band.

Conclusion

About 60% of the assigned values by the participants' consensus appear to be significantly biased. The observed bias is confounded between the field laboratory method and the commercially available IVD kits used for the measurements hence the estimated bias may be method dependent. Getting unbiased creatinine measurements is a clear opportunity for improvement with straight implications for the population's health, either for diagnostics or treatment. The uncertainty of the participants cannot be explained by the argument of biological variability[9]. Although this improvement opportunity was intended for the field laboratories it may be fixed by the IVD kit producers. The use of CRM is justified for PT schemes and traceability recommendation of JCTLM to IVD kit producers in order to warrant traceability of their calibrants.

Acknowledgement

Thanks to the technical staff of the Division of Materials Metrology, whom experiments and professionalism made possible this data: Esther Castro Galván, Miryan Balderas Escamilla, Laura Regalado Contreras, Judith Rivera Mellado, Gabriela Salazar Briones, Mauricio Maldonado Torres, Marco Antonio Ávila Calderón, Melina Pérez Urquiza. Special thanks to the referee, whom critical feedback improved this work, Dr. Yoshito Mitani and Dr. Mariana Arce-Osuna for their enthusiastic support and critical feedback.

References

[1] Mitani & Pérez, 2002. Informe de resultados obtenidos en la prueba de aptitud técnica 630-IL-1002b para labporatorios quimico-clinicos que realicen mediciones de glucosa, colesterol, creatinina y calcio en suero. CENAM. Querétaro, México.

[2] Mitani & Pérez, 2004. Informe del estudio comparativo 630-Q023-1003 cuantificación de glucosa, creatinina, colesterol, ácido úrico y urea en suero humano. CENAM. Querétaro, México.

[3] Mitani & Pérez, 2006. Informe de la prueba de aptitud técnica 630-Q023-0054-PA Suero Humano. CENAM. Querétero, México.

[4] BIPM-JCGM/WG2, 2006. International Vocabulary of Metrology – Basic and General Concepts and Associated Terms.

[5] Huber, 1981. Robust statistics, John Wiley. NY.

[6] Efron, 1979. Bootstrap methods: Another look at the Jackknife. The Annals of Statistics, Vol 7, No 1, pp 1-26.

[7] ISO 13526, 2005. Statistical methods for use in proficiency testing by laboratory comparisons.

[8] BIPM-JCGM-100, 2008. Evaluation of measurement data – Guide to the expression of uncertainty in measurement.

[9] Harris et al, 1970. Biological and Analytic Components of Variation in Long-Term Studies of Serum Constituents in Normal Subjects. II. Estimating Biological Components of Variation. Clin. Chem; 16, 1022-1027.

[10] Thompson & Ellison, 2011. Dark uncertainty. Accred.Qual.Assur.16, pp 383-387.

[11] Duewer et al, 1999. Micronutrients Measurement Quality Assurance Program: Helping Participants Use Interlaboratory Comparison Exercise Results to Improve their Log-Term measurement performance. Anal.Chem.71, pp 1870-1878.

[12] Vangel & Rukhin. (1999). Maximun Likelihood Analysis for Heteroscedastic One-Way Random Effects ANOVA in Interlaboratory Studies. Biometrics, Vol 55, No 1, 129-136.

[13] NOM-064-SSA1-1993. Norma Oficial Mexicana que establece las especificaciones sanitarias de los equipos de reactivos utilizados para diagnóstico.