Analytical performance specifications in clinical chemistry: the holy grail?

Wytze P. Oosterhuis

Department of Clinical Chemistry and Haematology, Zuyderland Medical Center, Heerlen, The Netherlands

Correspondence to: Wytze P. Oosterhuis. Department of Clinical Chemistry and Haematology, Zuyderland Medical Center, Henri Dunantstraat 5, 6419 PC Heerlen, The Netherlands. Email: w.oosterhuis@zuyderland.nl.

Abstract: The conference in Milan in 2014 resulted in a consensus statement on quality specifications. An EFLM Task and Finish Group was started to deal with issues that remained unresolved and needed further development including total analytical error methods. Most importantly were questions concerning measurement uncertainty (MU) and flaws in the model for calculating the permissible (or allowable) total error (TE). This is related to performance specifications based on biological variation, the second model of the Milan consensus. Performance specifications are closely linked to the work of Westgard and the concept of TE, that depends on the combined effect of the random and systematic errors of the method which is compared to a permissible TE that is in most cases based on biological variation. Where the first models for performance specifications did not include bias, later models did. The concept of bias is however complicated, both in the estimation of the value, and in the integration of bias in a mathematical model. In many fields outside clinical chemistry, the error concept has been abandoned in favour of the MU model. Bias is commonly excluded from this model, where only the uncertainty of bias is taken into account. There is no general consensus on these issues, and challenges remain to integrate the different concepts, both for the definition of performance, of performance specifications as for quality control procedures.

Keywords: Performance specifications; quality; total error (TE); measurement uncertainty (MU)

Received: 01 September 2017; Accepted: 04 September 2017; Published: 26 September 2017.
doi: 10.21037/jlpm.2017.09.02
View this article at: http://dx.doi.org/10.21037/jlpm.2017.09.02

Introduction

In 2014 a conference on quality specifications in Milan resulted in a consensus statement (1) that was based on the previous Stockholm consensus of 1999 (2). There were some developments acting as incentives for the Milan conference. Most important were the requirements of ISO 17025 and 15189 standards that laboratories should routinely provide the measurement uncertainty (MU) of the results, and recent comments on the total error (TE) theory stating that it contained flaws (3). It was acknowledged during the Milan conference that many issues including total analytical error methods remain unresolved and needed further development. An EFLM Task and Finish Group on Total Error was established for that purpose.

Performance specifications set limits for a test to establish whether this test is acceptable for routine use. Different specifications are needed for diagnosis and monitoring. This also includes optimum goals that may be unachievable by current state-of-the-art procedures. It is vital to obtain a tool for defining the ideal specifications without the influence of the state-of-the-art analytical quality, as this will set a goal for manufacturers.

One might wonder why more than 50 years after Tonks' proposal—that the maximum permissible imprecision should be less than one quarter of the reference range—the discussion of performance specifications still has not ended. It is only part of the challenge to calculate the uncertainty of a test, be it by using a model based on the MU paradigm, or a model based on the TE theory (as the TE can also be considered as a measure of uncertainty of the test result). However, a bigger challenge is to define the limits of

© Journal of Laboratory and Precision Medicine. All rights reserved.

acceptability of the analytical quality, or the performance specifications. Some of the issues that remain subject for dispute in this field are reviewed and commented here.

The Milan performance specifications

Following the Milan conference, criteria were formulated to assign measurands to appropriate models (or principles) for analytical performance specifications (4). The preferred performance specification model is based on clinical needs. This seems to be the most logical and obvious principle. It should be acknowledged, however, that this model can only be applied for a few measurands that have clearly defined decision levels (e.g., HbA1c, cholesterol). The second principle is based on models that use (components of) biological variation and can be applied for measurands that are in steady state or can be "transformed" to a steady state situation in biological fluids (4). This principle can—and is—most frequently applied. A third principle is based on the state-of-the-art and can be applied in cases where models 1 and 2 cannot be used (e.g., tests in urine samples).

TE

Performance specifications are closely linked to the work of Westgard. In 1974 Westgard, Carey, and Wold (5) introduced the concept of Total Analytical Error to provide a quantitative measure for the acceptability of analytical performance. Where reference laboratories estimate imprecision and bias separately by replicate measurements, clinical laboratories routinely measure patient- and quality assurance samples only once. According to this concept, the total analytical error in these circumstances depends on the combined effect of the random and systematic errors of the method which is compared to a defined allowable- or permissible TE (TE allowable TEa, or pTAE). The analytical error defines the maximum error for patient results that a single result can show with a certain probability, in most cases 95%. The analytical error thus estimates the limits of an interval around the true value where measured analytical results can be found with a defined probability. This model further assumes that the difference between the patients' result and the true value can be estimated primarily from results from proficiency testing or from internal quality assurance.

Estimating the TE of a test is just one side of the model. The other side is to set the limits to the total analytical error that can be tolerated in a test result without compromising its medical usefulness. In the TE model the linear model is used, that can be regarded as the conventional model.

MU

Uncertainty methods originated in physical measurements and chemistry (6,7). Laboratory medicine is still struggling to adapt to this long tradition established by physical metrology laboratories (8). The main differences between TE and MU are related to the concept of a true value and the related error concept, and the concept of bias and how to deal with bias.

In the TE model we look from the perspective of the true value, with the TE as an estimate of the difference between measurement result and true value. In the MU model we look from the perspective of the measurement result, with a confidence interval representing the total uncertainty of the measurement. The measurement result is related through calibration with the reference value, that represents a best estimate of the purely hypothetical "true" value.

The MU concept also assumes that if the bias of a procedure is known, then steps are to be taken to minimize it, e.g., by re-calibration. However, because the bias value cannot be known exactly, an uncertainty will be associated with such a correction. Thus, in the MU concept, a measurement result can comprise two uncertainties: the uncertainty due to imprecision, and the uncertainty associated with the bias correction. The uncertainties that act on the measurement result are combined to one MU statistic. Bias as it is included in the TE model thus is contradictory to the MU concept: when bias is known, it should be corrected.

Models for performance specifications

The approach of relating analytical performance to the biological variation was very much inspired by the ideas of Tonks (9). He empirically stated that the permissible analytical variation should not exceed "one quarter" of the reference interval. This can more or less be considered as a performance specification based on biological variation as the reference interval is mainly determined by the latter. Since Tonks many alternative models have been suggested for calculating performance specifications based on these principles. Cotlove *et al.* (10) proposed that analytical goals should not be based on the reference interval, because this reflects both the analytical and the biological variation and represents circular reasoning. He suggested to only include

© Journal of Laboratory and Precision Medicine. All rights reserved.

the total biological standard deviation (SD_{biol}) composed of the within (SD_I) and the between-subject or group standard deviation (SD_G): $SD_a < 0.5SD_{biol}$ with $SD_{biol} = (SD_I^2 + SD_G^2)^{0.5}$.

Monitoring vs. diagnosis

For monitoring Cotlove and Harris proposed to include only the within subject variation: $SD_a < 0.5SD_I$ (10). When a test is being applied to confirm or rule out a particular diagnosis, they stated that a population-based reference range will be used in the interpretation of the test result, based on the total biological variation. On the other hand, when the test is one of a series over time, as in monitoring the patient's status, only intra-individual biological variation is relevant. It follows that, since the relevant biological variation is smaller in the monitoring situation, the size of the analytical variation of the test will be more critical here (11). This reasoning was accepted at the 1976 College of American Pathologists (CAP) Conference (12). It should be noted that the question of bias was not addressed.

Combining bias and imprecision in the calculation of performance specifications

Harris (11) expanded his original work (13) so that bias and imprecision were both taken into account. The original equation for monitoring: $SD_a < 0.5SD_I$ was changed into $(SD_a^2 + Bias^2)^{0.5} < 0.5SD_I$.

In the same year Gowans *et al.* published a model based on reference intervals, that formed the basis for many other studies (14). Like the model of Harris, it derived performance specifications for different combinations of bias and imprecision. With the transferability of reference intervals between laboratories as starting point, the permissible bias and imprecision are calculated based on the premise that the reference interval limits will remain valid with a maximum of false positives exceeding the reference limit of 4.6% (instead of the usual 2.5%). Like the model of Harris, the resulting relationship between maximum permissible bias and imprecision is curved. Both extremes of this curve are interesting: with bias =0, maximum permissible $CV_a = 0,6CV_{biol}$; with $CV_a = 0$ (a hypothetical value) bias $= 0.25CV_{biol}$. Although these values are mutually exclusive, these values are used together in conventional models from permissible analytical performance. Note that the model is based on reference values and total biological variation, and should thus be applied for diagnosis and not for monitoring.

The "conventional" linear model

The model that is used most frequently is the model proposed by Hyltoft Petersen and Fraser in 1993. Bias and imprecision specifications are combined to set quality limits based on biological variation: (15,16):

$$pTAE = 0.25\ (CV_I^2 + CV_G^2)^{0.5} + 1.65\ (0.5CV_I)\ [1]$$

Note that the linear combination of terms for bias and imprecision follows the same reasoning as the TE model. The performance specification was proposed for proficiency testing, but has been extensively used to define specifications for other purposes, e.g., in listings of permissible TE (17). The term for bias is that according to Gowans' model combined with the generally accepted maximum imprecision of $0.5CV_I$ (with a coverage factor of 1.65, corresponding to P=0.95, one-sided). This expression shows a linear relationship between bias and imprecision, and assumes a fixed value for pTAE for all combinations of maximum bias and imprecision.

Fraser (18) adapted this expression, because some measurands are subject to tight homeostatic control (e.g., electrolytes), leading to unrealistic performance specifications. Three quality levels are used for imprecision and bias: optimum ($CV_a \leq 0.25CV_I$, Bias $\leq 0.125CV_{biol}$), desirable ($CV_a \leq 0.5CV_I$, Bias $\leq 0.25CV_{biol}$) and minimum ($CV_a \leq 0.75CV_I$, Bias $\leq 0.375CV_{biol}$).

Other models and adaptations

The model of Gowans and other models based on reference limits assume these limits to be based on biological variation alone. This is clearly an oversimplification in some measurants. Oosterhuis and Sandberg (19) adapted the model of Gowans *et al.* (14) to include the influence of analytical variation on the reference interval. Even inclusion of analytical variation, however, might lead to an underestimation of the actual reference interval limits e.g., due to pre-analytical variation. As an alternative the actual reference interval limits can be used as starting point in the model (20).

It should be noted that in most distributions CV_G and CV_I are log-Gaussian, as are most reference ranges (21). Performance specifications derived from biological data should ideally be based on this log-Gaussian distribution (20). Other models combining bias and imprecision for calculation of pTAE might also be considered e.g., (22-25). Most importantly, for monitoring models have been developed based on reference change values. As most tests are used for monitoring, these should be considered to be the dominant models (26).

Discussion

At the Milan conference in 2014 the performance specifications were re-formulated based on the Stockholm consensus: the idea of a hierarchy was changed to models for specifications that would fit best for the measurand. The first model is based on the effect of analytical performance specifications on clinical outcome. This is the model of choice for measurands that have a central role in the decision-making of a specific disease or clinical situation and where clear decision limits are established. Total cholesterol, glucose, HbA1c, serum albumin and cardiac troponins represent examples. The second model is based on components of biological variation and should be applied to measurands where the first model is not applicable. The measurand should be in a steady state as in homeostatic control. The last model is based on state-of-the-art of the measurement, and should be used for all the measurands that cannot be included in models 1 or 2 (4).

Although the Milan consensus speaks of models, the proposal has a general character and specific models were outside the scope of the consensus. Incentives for the Milan conference had been both questions concerning the validity of the estimation of the performance specifications based on the conventional TE model, and the role of MU in the field of clinical chemistry. Concerning these questions that relate mainly to the second model, an EFLM Task and Finish Group was started.

At this point one might wonder why the analytical performance specifications are still subject to discussion. It is more than 50 years after the proposal of Tonks and also more than 40 years after the concept of Cotlove was accepted by the 1976 conference of the CAP, accepting performance specifications both for diagnosis as for monitoring (12). However, the Stockholm conference followed in 1999 and the Milan conference in 2014. What are the problems that are yet still discussed and that remain to be solved?

One important issue is that the validity of the conventional TE model is compromised by several flaws in the calculation of the permissible TE (pTAE) based on biological variation using expression [1] (17). There were several points of discussion related to pTAE that were addressed by the Task and Finish Group and pTAE was criticized for a number of reasons (27). Both maxima of permissible bias and imprecision are added to obtain pTAE, a pragmatic solution first proposed for the use in proficiency testing (3). However, the theoretical basis for this is lacking. Two maximum permissible errors are added, derived under the mutually exclusive conditions. The sum will allow an increase of the percentage of test results exceeding the predefined limits.

Another flaw in the model concerns the maximum permissible bias that was derived as $0.25CV_{biol}$ or $0.25\ (CV_I^2 + CV_G^2)^{0.5}$. It should be noted, however, that in the conventional model this bias term is applied in the case of *monitoring* although this expression had been derived by Gowans (14) from a population based reference range model and only applies to *diagnosis*. As an alternative, a model based on a reference change value model was developed that is only based on CV_I and not on CV_G (26).

Finally it has been argued that the condition $CV_a < 0.5CV_I$ relates to performance specifications that—according to the TE model—will lead to a sigma metric below 3 that cannot be maintained by internal quality assurance (27).

If these problems are clear and when there is agreement on these flaws, why could the TE model not be corrected? This proves not to be easy, and the Task and Finish group succeeded in the groundwork for further developments. What are the problems that should be resolved?

Definition of bias and imprecision

The first authors (9,10,13) did not include bias in their quality specifications models. Bias proves to be a difficult concept. GUM defines bias as any error that is reproducible, without defining the time frame. The bias concept does not fit well in our continuous 24/7 work flow. Bias that might be reproducible during a short period (one day, one week) might change over longer periods. So, both imprecision and bias are not independent of the choice of the period these are measured. One can distinguish between short-term bias (e.g., within day, one shift) and long-term bias (e.g., during several weeks or months): many effects causing short-term bias, e.g., re-calibrations may be seen as bias within this short time frame, but may be indistinguishable from random effects when variation is observed over a longer time period. When uncorrected, many short-term bias components increasingly contribute to the random error component of the MU. These effects will make any definition of bias and imprecision in part arbitrary.

© Journal of Laboratory and Precision Medicine. All rights reserved.

Performance specifications and quality control limits

It might seem logical to apply the same performance limits for analytical performance as the limits for internal quality assurance. However, there are reasons to set different limits: quality assurance applies rules to achieve a high probability of error detection and at the same time a low probability of false rejection, in most cases based on a singleton measurement result. Quality assurance limits will generally be stricter—e.g., by 1.65 SD_a—than performance limits in order to maintain the performance goals and assure that—within a pre-defined probability—that these goals are achieved. It might, however, seem a paradox that when the results of quality assurance measurements are within the limits, in retrospect these results will be well within the pTAE limits with a wide margin (28).

Six Sigma and quality control perspectives

Quality specifications can be derived from biological variation or from other specifications based on clinical needs. We can measure the precision of the test and calculate the sigma metric to express the quality on the sigma scale with 6 as very good and 3 as just sufficient quality to maintain with quality control procedures. Based on the sigma metric the appropriate quality control procedures can be developed. Two different points of view are in play here. On the one hand we can focus completely on the clinical needs, with the technical aspects of the test as secondary to these specifications. As the sigma metric will be derived from the clinical needs, this represents the Six Sigma perspective. Following this line of thought, this will lead to very relaxed quality control rules in tests that have a high sigma score, and strict rules with low sigma scores.

On the other hand we could focus completely on the technical aspects of the test without taking clinical aspects into account. The control limits in this case are derived solely on the imprecision of the test (e.g., ±3SDa).

The difference between these two viewpoints is this: when we are only interested in the clinical perspective (that includes biological variation), we will not be interested in changes signaling that the measuring system is more or less out of control, and quality control results that are outside 3SD_a limits do mean just that. However, even if the system is out of control from the technical point of view, the test results could be within clinical needs. Looking from the clinical perspective, these abnormal results have no meaning. Are we to know the system is out of control and should we act upon this, even if there is no clinical need?

Performance specifications based on biological variation or reference values?

A commonly raised critical comment on the use of reference ranges as the basis of performance specifications is that reference ranges themselves depend on analytical performance, leading to a circular reasoning. Although the contribution of the analytical variation to the total variation will in many cases be small (29), it is a simplification to assume that reference ranges are only determined by biological variation as has been done in many models. As an alternative the analytical variation has been included in an adaptation of the model of Gowans (14). In this model the analytical variation is included as it was when the reference ranges were determined or confirmed (19).

Combining MU and TE models

MU and TE represent different paradigms in metrology. TE is based on the error concept; in order to calculate the error, however, the true value or at least a reference value should be known. The concept of a “true value” that once was the cornerstone in metrology has been abandoned by GUM (30). In MU we only have the concept of the uncertainty of the measurement result, with a value that—through calibration—can be traced back to a standard of a higher order. Another difference is the treatment of bias: in the MU paradigm, bias should be corrected when known.

The Task and Finish Group concluded that the MU model fits well for patients’ test results, while the TE model can be applied for quality control purposes. The main reason for this is, that in quality control there is a reference or target value that allows us to calculate the error. However, in patients there is no reference value and the result could be expressed with an estimate of the uncertainty. This uncertainty is based on all sources of variation, including preanalytical factors and biological variation.

This still leaves open what error model to be used in quality control and how to determine quality limits. The bias concept still remains a problem, and we might even abandon the bias concept altogether and assume all forms of error (deviation from the reference value) as short- or long term imprecision. We should be able to include in a model the maximum permissible difference between analysers performing the same test within one laboratory organisation.

Conclusions

The Task and Finish group agreed on certain issues: flaws in conventional model, and the application of MU and TE models. However, there are still challenges: how to develop and agree upon an integrated system of MU, permissible analytical error, defining and dealing with bias and how to develop quality control rules possibly within the Six Sigma model. It has been stated "all models are wrong, but some models are useful" (31). The practice of the clinical laboratory is such, that it is impossible to describe performance specifications in a mathematically perfect model, and all models will be based on assumptions and can only approach complex reality. The challenge is to reach consensus on a model that is both useful and as less flawed as possible.

Acknowledgements

With thanks to the members of the Task and Finish Group on Total Error for the fruitful discussions.

Footnote

Conflicts of Interest: The author has no conflicts of interest to declare.

References

1. Sandberg S, Fraser CG, Horvath AR, et al. Defining analytical performance specifications: Consensus Statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. Clinical chemistry and laboratory medicine 2015;53:833-5.
2. Kenny D, Fraser CG, Petersen PH, et al. Consensus agreement. Scand J Clin Lab Inv. 1999;59:585.
3. Oosterhuis WP. Gross overestimation of total allowable error based on biological variation. Clin Chem 2011;57:1334-6.
4. Ceriotti F, Fernandez-Calle P, Klee GG, et al. Criteria for assigning laboratory measurands to models for analytical performance specifications defined in the 1st EFLM Strategic Conference. Clin Chem Lab Med 2017;55:189-94.
5. Westgard JO, Carey RN, Wold S. Criteria for judging precision and accuracy in method development and evaluation. Clin Chem 1974;20:825-33.
6. Page CH, Vigoureux P. The International Bureau of Weights and Measures 1875-1975. National Bureau of Standards; 1975.
7. Williams A. What can we learn from traceability in physical measurements? Accredit Qual Assur 2000;5:414-7.
8. Williams A. Traceability and uncertainty - A comparison of their application in chemical and physical measurement. Accredit Qual Assur 2001;6:73-5.
9. Tonks DB. Quality Control Systems in Clinical Chemistry Laboratories. Postgrad Med 1963;34:A58-A70.
10. Cotlove E, Harris EK, Williams GZ. Components of variation in long term studies of serum constituents in normal subjects III. Physiological and medical implications. Clin Chem 1970;16:1028-32.
11. Harris EK. Proposed goals for analytical precision and accuracy in single point testing. Arch Pathol Lab Med 1988;112:416-20.
12. Proceedings of the Subcommittee on Analytical Goals in Clinical Chemistry, World Association of Societies of Pathology, Ciba Foundation, London, England (U. K.), April 25-28, 1978. Am J Clin Pathol 1979;71:624-30.
13. Harris EK. Statistical principles underlying analytical goal-setting in clinical chemistry. Am J Clin Pathol 1979;72:374-82.
14. Gowans EM, Hyltoft Peteresen P, Blaabjerg O, et al. Analytical goals for the acceptance of common reference intervals for laboratories throughout a geographical area. Scand J Clin Lab Invest 1988;48:757-64.
15. Fraser CG, Petersen PH. Quality Goals in External Quality Assessment Are Best Based on Biology. Scand J Clin Lab Invest Suppl 1993;212:8-9.
16. Fraser CG, Petersen PH. Desirable standards for laboratory tests if they are to fulfill medical needs. Clin Chem 1993;39:1447-53; discussion 1453-5.
17. Ricos C, Alvarez V, Cava F, et al. Desirable Specifications for Total Error, Imprecision, and Bias, derived from intra- and inter-individual biologic variation. Available online: http://www.westgard.com/biodatabase1.htm. (Assessed 31 Aug. 2017).
18. Fraser CG, Hyltoft Petersen P, Libeer JC, et al. Proposals for setting generally applicable quality goals solely based on biology. Ann Clin Biochem 1997;34:8-12.
19. Oosterhuis WP, Sandberg S. Proposal for the modification of the conventional model for establishing performance specifications. Clin Chem Lab Med 2015;53:925-37.
20. Haeckel R, Wosniok W, Gurr E, et al. Permissible limits for uncertainty of measurement in laboratory medicine.

Clin Chem Lab Med 2015;53:1161-71.
21. Hyltoft Petersen P, Blaabjerg O, Andersen M, et al. Graphical interpretation of confidence curves in rankit plots. Clin Chem Lab Med 2004;42:715-24.
22. Åsberg A, Odsæter IH, Carlsen SM, et al. Using the likelihood ratio to evaluate allowable total error - an example with glycated hemoglobin (HbA(1c). Clin Chem Lab Med 2015;53:1459-64.
23. Oddoze C, Lombard E, Portugal H. Stability study of 81 analytes in human whole blood, in serum and in plasma. Clin Biochem 2012;45:464-9.
24. Macdonald R. Quality assessment of quantitative analytical results in laboratory medicine by root mean square of measurement deviation. J Lab Med 2006;30:111-7.
25. White GH, Farrance I; AACB Uncertainty of Measurement Working Group. Uncertainty of measurement in quantitative medical testing: a laboratory implementation guide. Clin Biochem Rev 2004;25:S1-24.
26. Larsen ML, Fraser CG, Petersen PH. A comparison of analytical goals for haemoglobin HbA1c assays derived using different models. Ann Clin Biochem 1991;28:272-8.
27. Oosterhuis WP, Bayat H, Armbruster D, et al. The use of error and uncertainty methods in the medical laboratory. Clin Chem Lab Med 2017. [Epub ahead of print].
28. Ceriotti F, Brugnoni D, Mattioli S. How to define a significant deviation from the expected internal quality control result Clin Chem Lab Med 2015; 53:913-8.
29. Haeckel R, Wosniok W, Gurr E, et al. Supplements to a recent proposal for permissible uncertainty of measurements in laboratory medicine. Laboratoriums Medizin 2016;40:141-5.
30. JCGM. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. JCGM 100:2008, GUM 1995 with minor corrections. Available online: http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf. Joint Committee for Guides in Metrology; 2008. (Assessed 31 Aug. 2017).
31. Box GEP. Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN. editors. Robustness in Statistics, Academic Press, 1979:201-36.

Cite this article as: Oosterhuis WP. Analytical performance specifications in clinical chemistry: the holy grail? J Lab Precis Med 2017. doi: 10.21037/jlpm.2017.09.02;2:78