Opinion Paper

Geir Thue and Sverre Sandberg*

Analytical performance specifications based on how clinicians use laboratory tests. Experiences from a post-analytical external quality assessment programme

Abstract: Analytical performance specifications can be based on three different models: the effect of analytical performance on clinical outcome, based on components of biological variation of the measurand or based on state-of-the-art. Models 1 and 3 may to some degree be combined by using case histories presented to a large number of clinicians. The Norwegian Quality Improvement of Primary Care Laboratories (Noklus) has integrated vignettes in its external quality assessment programme since 1991, focusing on typical clinical situations in primary care. Haemoglobin, erythrocyte sedimentation rate (ESR), HbA$_{1c}$, glucose, u-albumin, creatinine/estimated glomerular filtration rate (eGFR), and Internationl Normalised Ratio (INR) have been evaluated focusing on critical differences in test results, i.e., a change from a previous result that will generate an "action" such as a change in treatment or follow-up of the patient. These critical differences, stated by physicians, can translate into reference change values (RCVs) and assumed analytical performance can be calculated. In general, assessments of RCVs and therefore performance specifications vary both within and between groups of doctors, but with no or minor differences regarding specialisation, age or sex of the general practitioner. In some instances state-of-the-art analytical performance

could not meet clinical demands using 95% confidence, whereas clinical demands were met using 80% confidence in nearly all instances. RCVs from vignettes should probably not be used on their own as a basis for setting analytical performance specifications, since clinicians seem "uninformed" regarding important principles. They could rather be used as a background for focus groups of "informed" physicians in discussions of performance specifications tailored to "typical" clinical situations.

Keywords: analytical performance specifications; case histories; EQA; erythrocyte sedimentation rate (ESR); estimated glomerular filtration rate (eGFR); glucose; haemoglobin; HbA$_{1c}$; International Normalised Ratio (INR); outcome; post-analytical; quality; quality specifications; tests; u-albumin.

Introduction

Analytical performance specifications can in principle be set in three different ways [1], i.e., based on the effect of analytical performance on clinical outcome, based on components of biological variation of the measurand or based on state-of-the-art. One way of indirectly exploring the probable effect on the outcome is to explore how physicians or clinical experts use the tests or would like to use the tests. Both biological variation and state-of-the-art will be expressed with some numerical "uncertainty". Clinical opinions will also be at variance, both due to differences in perception of the clinical situation at hand, but also due to differences regarding knowledge and experience, e.g., with consequences of deviant results. Still, this method is attractive, since analytical performance could be directly tailored to clinical use.

In principle there are two main methods to explore clinical opinion, attaining information from many

*Corresponding author: Sverre Sandberg, Norwegian Quality Improvement of Primary Care Laboratories, Department of Global Public Health and Primary Care, University of Bergen, Boks 6165, 5892 Bergen, Norway; and Laboratory of Clinical Biochemistry, Haukeland University Hospital, Bergen, Norway, E-mail: sverre.sandberg@isf.uib.no
Geir Thue: Norwegian Quality Improvement of Primary Care Laboratories, Department of Global Public Health and Primary Care, University of Bergen, Norway

physicians: 1) use of journal notes to see how physicians use the laboratory test in real life, i.e., what changes in test results that generate a change in the treatment of the patient; and 2) looking at the intended use by for example distributing case histories to simulate the real life situation. The latter has the advantages of standardisation and is not so prone to differences in perception. In vignettes, clinicians are often asked what "critical difference" in a test result that will generate an "action", i.e., a change in treatment or follow-up of the patient. This critical difference can translate into the term reference change value (RCV) that is most often used in laboratory medicine. Dependent on the question asked, this RCV can comprise pre-analytical variation, imprecision, within-subject variation, and bias. The first ones to use this method were Elion-Gerritzen [2] and Skendzel [3] who included questions on RCVs in their studies, and thereafter calculated the corresponding analytical imprecision. They did, however, not take into account the within-subject variation or other possible sources of variation.

The Norwegian Quality Improvement of Primary Care Laboratories (Noklus) has performed analytical external quality assessment (EQA) in primary care for many years, and has strived to integrate interpretation of laboratory results in the program using questionnaires, i.e., in a post-analytical EQA. This method should be especially useful in primary care since decisions in many instances are less complex and based on relatively few data. Some of the studies were carried out in cooperation with the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) and the European Organisation of External Quality Assurance Providers in Laboratory Medicine (EQALM). The purpose of this paper is to sum up results, focus on experiences, and to evaluate how the clinical vignette method could be useful to the laboratory-clinician interaction, especially with regard to what performance specifications physicians assume when interpreting laboratory results.

Materials and methods

During 1991–2012 Noklus published a series of papers on post-analytical quality assessment addressing how physicians used and interpreted laboratory tests both in Norway and internationally, with mainly European data [4–13]. All but one study used case histories with feedback reports as the research/educational instrument.

The formulation of case histories involved six important elements: 1) several or a specific and typical situation in which the analyte was important for monitoring or diagnosis were identified using journal notes and clinical experience; 2) results of other tests should not be necessary in the situation described; 3) only one decision was elucidated, so that answers were not conditional on an earlier decision; 4) the clinical evaluation needed should be frequently encountered; 5) the histories should be short, with only essential information, and 6) the writing should be partly "conversational" to make the clinical situation recognisable. The case histories were piloted by general practitioners (GPs) to ensure face validity, and sometimes reviewed by consultants in relevant specialities and academic staff. Comments were invited in the questionnaires as part of the validity evaluation. In one study, laboratory results from hospital databases were used, and the GP thus evaluated a real patient's laboratory result, using actual journal information. Most case histories were presented to GPs, but on some occasions to other types of specialists as well as specialists in laboratory medicine.

Part of the questionnaire was always denoted to clinically important ("critical") differences (RCVs), i.e., mostly changes between two laboratory results, or between a

---

A 45-year-old obese woman with five children. She had type 2 diabetes diagnosed 4 years ago, for which she is treated with oral antidiabetics. She is also treated with an ACE-inhibitor for hypertension (BP140/100 mm Hg before treatment). You are her physician. Her life style is hectic and she pays little attention to diet and exercise. She monitors her blood glucose a few times a month and the results vary between 7 and 16 mmol/L. You are not sure that her monitoring is performed correctly.

At the present consultation her HbA_{1c} is 9.1%
You do what you find appropriate.
In your opinion, what should the HbA_{1c} test result be at the next consultation for the value to indicate:
A. Better diabetes control: HbA_{1c} must have decreased to at least   __,__ %
B. Worse diabetes control: HbA_{1c} must have increased to at least   __,__ %

Figure 1: Example of a case history presented to general practitioners in six countries [8], eliciting information on the reference change value of HbA_{1c} in follow-up of a diabetes patient.

laboratory result and a reference limit large enough to lead to actions not otherwise undertaken (e.g., change of medication). A preset laboratory result, or the office laboratory EQA result was used as the initial value, and an example of a case history is stated in Figure 1. Time span between results varied, from days to weeks or months, and both capillary and venous samples were presumed, thus focusing on analytical variation, lot-to-lot variation, and bias. However, within-subject biological variation is also important when evaluating RCVs, depending on the magnitude of biological versus analytical variation. The RCV can be calculated when the within-subject variation (CV_i) and the analytical variation (CV_a) as well as the probability, z-score, is known [5].

$$RCV = bias + z\sqrt{2} \cdot \sqrt{CV_i^2 + CV_a^2}$$

When clinicians are asked to state a RCV, the formula can be rearranged to calculate the CV_a:

$$CV_a = \sqrt{\left(\left(\frac{RCV - bias}{z} \right) \cdot \sqrt{2} + CV_i^2 \right)}$$

where CV_a is the analytical imprecision, CV_i is the within subject biological variation, z is the probability or confidence with which it is assumed that the clinician takes an action based on the difference between the laboratory results.

Assuming that the bias is zero or known, the CV_a presumed by clinicians as necessary to fulfil clinical needs can be calculated (unless impossible when the "square root" expression is negative), and compared to hospital laboratory or point-of-care analytical imprecision. Most often the CV_a was based on the median RCV, denoting desirable performance. Within-subject biological variation, CV_i was obtained from the literature. One side test z-values referred to 95% of 80% confidence, and the square root of 2 was omitted when a change from a reference limit was to be stated as denoted in [5].

Results

In general, assessments of RCVs and therefore the performance specifications vary substantially both within and between groups of doctors, but with no or minor differences regarding GP age, sex or whether or not the GP was a specialist in general practice. The same holds true when including other specialists. Stated RCVs in several settings were not congruent regarding the size of increase

or decrease from an initial value. In some instances state-of-the-art analytical performance could not meet clinical demands using 95% confidence, whereas clinical demands were met using 80% confidence in nearly all instances. An overview of the studies is set out in Table 1, and an example of a vignette is presented in Figure 1.

Hemoglobin was first assessed in 1991 using several case histories [4], and then again in 1999 as part of an EQAS in Norway [5]. A 62-year-old male had undergone hemicolectomy for cancer 2 years ago, and a fall in hemoglobin necessary to take action should be stated by the GPs. Both desirable, optimum and minimum performance specifications based on 50, 25 and 75 percentiles of responses, respectively, were attainable with instruments used in general practice [intra- and inter-office CV_a of 1.4% and 2.4%, respectively (Noklus, unpublished data)].

Evaluation of the erythrocyte sedimentation rate (ESR) test was done by presenting 12 case histories to Norwegian GPs [6]. In principle, for many GPs the action value increased as the given ESR increased, whereas others reacted on a constant change in ESR, or the change necessary to take action was highly dependent on the clinical situation. In many instances, half the GPs reacted on a change of 10 mm/h or less, a change that can be explained by biological and analytical variation.

HbA_{1c} was first assessed in 1997 in a national study in Norway [7], and then in an international survey (also including blood glucose) to GPs in five other countries during 2001–2002 [8]. The case history told of a 45-year-old woman with type 2 diabetes and poor metabolic control, focusing on HbA_{1c} results in monitoring (Figure 1). Baseline HbA_{1c} was either a given value, or a (similar) result obtained in an accompanying EQAS. Judgements were irrespective of whether the GPs analysed HbA_{1c} in their practice laboratory or not. Second, the pattern of judgement was similar between countries, with lower changes considered true when HbA_{1c} increased (from 9.1%) than when it decreased – but with considerable variations among GPs. Calculated CV_a based on the median RCV was attainable (both for an increase and decrease of dosage), but only when using 80% confidence instead of the conventional 95%. Since the vignette accompanied an EQAS for HbA_{1c} in the Norwegian study, we were able to compare RCVs to office laboratory performance. We found that in practices with "poor" analytical quality on their office instrument, 65% of the GPs stated a RCV smaller than the deviation of the EQAS result (on their office instrument) from the target value (reference value). Thus the GPs generally think that the analytical quality is better than it is and that they are not aware of the analytical quality of their own laboratories. This

Table 1: Overview of studies regarding reference change values (RCVs) and analytical performance specifications for imprecision (CV_a) for some constituents.

Analyte, countries (ref)	No. of participants (response rate)	Initial value	Median RCV	Median CV_a (95% confidence)	Median CV_a (80% confidence)
Hb, Norway [4, 5]	1500 (50%)	145 g/L	15 g/L	3.9%	–
ESR, Norway [6]	206 (76%)	Low ESR	10 mm/h	NC	–
HbA$_{1c}$, 6 countries [7, 8]	1109 (16%–70%)	9.1%	+0.5%–0.9% (increase)[a] −0.6%–2.1% (decrease)[a]	1.5–NC 2.2–9.1	3.5–9.7 7.4–25.0
Glucose, 6 countries [7, 8]	2000 (16%–70%)	5.8 mmol/L (capillary)	0.5–0.7 mmol/L (increase)[a] 0.5–0.6 mmol/L (decrease)[a]	1.5–NC NC	8.1–12.0 8.1–10.0
u-albumin (ACR), 9 countries [9]	2078 (7%–43%)	15 mg/mmol	20%–87% (increase)[a] 33%–73% (decrease)[a]	– 	14–81 14–60
s-Creatinine/eGFR, Norway [10] (based on actual results in laboratory databases; median values for females and males given as initial values)	210 (60%)	Creatinine 95 µmol/L females 124 µmol/L males eGFR 52–51 mL/min/1.73 m²	Improvement in renal function: creatinine 14 µmol/L – 12% eGFR 8 mL/min/1.73 m² – 17% deterioration: creatinine 20 µmol/L – 18% eGFR 8 mL/min/1.73 m² – 17%	2%	–
Creatinine/eGFR/ACR Norway/Netherland [11] Specialists in laboratory medicine	52 (52%)	Creat: 119 µmol/L eGFR 54 mL/min/1.73 m² ACR: 15 mg/mmol	−14% for improvement +14% for deterioration +18% improvement −13% deterioration −57% improvement +72% deterioration	2%	–
INR, Norway [12]	1547 (41%)	INR 3.3	INR 2.5 to increase dose INR 4.0 to decrease dose	4.1% NC	18.0 15.1
INR, 13 countries [13] GPs and specialists	3015 (8%–38%)	INR 2.3	INR 1.9 (1.8–2.0)[a] to incr. Dose INR 3.1 (3.0–3.4)[a] to decr. Dose	NC 12.7 (3.4–16.4)	–

NC, not possible to calculate. [a]Range between countries.

can cause misclassifications of the "patient's" clinical condition.

Blood glucose was assessed in six countries in 2001–2002, together with HbA_{1c} [8]. The focus was on case-finding, presenting a 64-year-old male with a capillary fasting glucose of 5.8 mmol/L, and asking the GP to state the next glucose result a few days later believed to be truly different from the initial result (RCV). Responses were rather similar comparing countries, but with substantial inter-GP variation. State-of-the-art or recommended analytical quality could not meet clinical demands at 95% confidence (Table 1).

Urine albumin was explored in 2006 using a case history depicting a male of 57 years with type 2 diabetes previously not tested for microalbuminuria [9]. A preset value of 15 mg albumin/mmol creatinine was given for the albumin/creatinine-ratio (ACR), and GPs in nine countries were asked to state a RCV for the ACR a year later. Other measurement units were available, but were not used in calculations of the assumed CV_a. RCVs (percentage change) were rather similar independent of reporting unit category, with attainable CV_a only for 80% confidence (Table 1).

Creatinine and estimated glomerular filtration rate (eGFR) were evaluated using data on real patients from laboratory databases [10]. Patients with a baseline eGFR of 30–59 mL/min/1.73m² were selected, one patient per GP, and the GP who had requested the analysis was then asked to state a RCV for creatinine and eGFR signaling an improvement or deterioration in renal function. Calculated CV_a's were attainable with 95% confidence. RCVs suggested were not influenced by the presence of albuminuria, multiple regression analysis did not explain differences between doctors, and relatively greater changes (percentage change) were stated for low eGFR values.

Creatinine, eGFR and ACR were also evaluated from the perspective of laboratory specialists advising GPs on test results, using two case histories and asking for RCVs from preset/given values [11]. Replies on Creatinine and eGFR were close to calculated values, with minor differences between doctors, and seemingly based on knowledge rather than intuition. For ACR suggestions were more at variance, and may reflect the diversity of data on biological variation in literature.

Finally, the International Normalised Ratio (INR) has been evaluated in two studies; first in Norway [12], and then internationally [13]. In the Norwegian survey, the case history described a 72-year-old male with mechanical heart valve prosthesis on stable (and strict) anticoagulation with warfarin. In the international study, a male of 76 years with atrial fibrillation and a stable INR was described (last result 2.3, range 2.0–2.8 during "the last months"). The anticoagulant treatment was according to medications used routinely in participating countries, since warfarin was not used in all countries. The GPs (in Norway), and the GPs or the secondary care specialists in the international study were to state INR-values necessary to change the treatment on a routine follow-up visit. Most GPs and specialists would change the dose at or right outside the therapeutic range, and in most countries responses were similar. No differences were found regarding speciality, use of dosing algorithms, or the availability of a point-of-care INR instrument. RCVs were attainable only when the assumed CV_a was calculated with 80% confidence.

Discussion

The main findings in these studies were the substantial variation between doctors when estimating clinically important RCVs. In many instances, the analytical quality presupposed by clinicians could not be met by state-of-the art technology using conventional 95% confidence, and seemingly the same quality was assumed irrespective of where the analysis was done. The variation was substantial and of clinical consequence.

The response rate was rather low in many of the studies, although not lower than in similar studies using vignettes [14, 15], whereas in some studies it is rather high and less prone to selection bias. In Norway participating GPs were representative for Norwegian GPs regarding sex and age distribution (statistics on file with the Norwegian Medical Association). Still, it is reasonable to assume that the most knowledgeable doctors respond, so variation is probably even larger.

Variation was comparable irrespective of analyte, e.g., variation for INR was similar for haemoglobin (data not shown), although evaluation of INR is much less dependent on clinical information and other laboratory results. Judging on the basis of a "real" result from an EQAS did not influence RCVs compared to a preset result "on paper" (for HbA_{1c}), the same goes for stating changes from a real laboratory result (for creatinine/eGFR) using journal information from consultations. For eGFR, it was even found that relatively greater changes were stated for low eGFR values, and changes were not related to the presence of albuminuria. These findings concerning real patients are clinically important, and support the assumption that the vignettes may be regarded as valid representations of doctors' practice behaviour. Indeed, alternatives to questionnaires for obtaining information on doctors'

evaluation of test results are difficult to imagine, since such information is not easily entangled from the complexities of the consultation. "Real life" evaluations would probably lead to even greater variation, since clinical situations will be perceived even more differently than when presented in questionnaires.

Judgements of laboratory results differed in several groups of primary care physicians and was in principle unchanged over time, but was much smaller for laboratory personnel asked to state RCVs with 95% confidence for creatinine and eGFR [11]. Thus, a more knowledgeable approach to RCVs is attainable. This finding probably means that primary care doctors in general are not familiar with the concept of biological variation and the probabilistic nature of laboratory information, and that analytical quality is always considered "acceptable".

In conclusion, clinical vignettes have a rationale although doctors' responses vary. First, vignettes may be the best way to monitor the clinical state-of-the-art regarding evaluation of analytical results. Using clinical vignettes to set analytical performance specifications is therefore in some way a combination of the Model 1 (clinical outcome) and 3 (state-of-the-art) since the performance specifications given by the clinicians will be highly influenced by the present analytical performance [1]. Moreover, the vignettes act as a means of two-way communication between clinicians and laboratory personnel, since analytical presumptions can be compared to actual performance, and the effects of biological variation and confidence are made explicit in recognisable and frequent clinical situations. This could be a stimulus to further information from the laboratory, or may even be used when informing physicians about a deviant result. However, results on test evaluation in vignettes should probably not be used on their own as a basis for setting performance specifications, since clinicians seem "uninformed" regarding important principles. They could rather be used as a background for focus groups of "informed" physicians in discussions of different levels of performance specifications tailored to frequent clinical situations.

References

1. Sandberg S, Fraser CG, Horvath AR, Jansen R, Jones G, Oosterhuis W, et al. Defining analytical performance specifications: consensus agreement from the 1st EFLM Strategic Conference. Clin Chem Lab Med 2015;53:833–5.
2. Elion-Gerritzen WE. Analytic precision in clinical chemistry and medical decisions. Am J Clin Pathol 1980;73:183–95.
3. Skendzel LP, Barnett RN, Platt R. Medically useful criteria for analytic performance of laboratory tests. Am J Clin Pathol 1985;83:200–5.
4. Thue G, Sandberg S, Fugelli P. Clinical assessment of haemoglobin values by general practitioners related to analytical and biological variation. Scand J Clin Lab Invest 1991;51:453–9.
5. Sandberg S, Thue G. Quality specifications derived from objective analyses based upon clinical needs. Scand J Clin Lab Invest 1999;59:531–4.
6. Thue G, Sandberg S, Fugelli P. The erythrocyte sedimentation rate in general practice: clinical assessment based on case histories. Scand J Clin Lab Invest 1994;54:291–300.
7. Skeie S, Thue G, Sandberg S. Use and interpretation of HbA1c testing in general practice. Implications for quality of care. Scand J Clin Lab Invest 2000;60:349–56.
8. Skeie S, Perich C, Ricos C, Araczki A, Horvath AR, Oosterhuis WP, et al. Postanalytical external quality assessment of blood glucose and hemoglobin A1c: an international survey. Clin Chem 2005;51:1145–53.
9. Aakre KM, Thue G, Subramaniam-Haavik S, Bukve T, Morris H, Müller M, et al. Postanalytical external quality assessment of urine albumin in primary health care: an international survey. Clin Chem 2008;54:1630–6.
10. Aakre KM, Thue G, Svarstad E, Skadberg Ø, Sandberg S. Laboratory investigation and follow-up of chronic kidney disease stage 3 in primary care. Clin Chim Acta 2011;412: 1138–42.
11. Aakre KM, Oosterhuis WP, Sandberg S. How do laboratory specialists advise clinicians concerning the use and interpretation of renal tests? Scand J Clin Lab Invest 2012;72:143–51.
12. Kristoffersen AH, Thue G, Sandberg S. Postanalytical external quality assessment of warfarin monitoring in primary healthcare. Clin Chem 2006;52:1871–8.
13. Kristoffersen AH, Thue G, Ajzner EB, Claes N, Horvath AR, Leonetti R, et al. Interpretation and management of INR results: a case history based survey in 13 countries. Thromb Res 2012;130:309–15.
14. Sood R, Sood A, Ghosh AK. Non-evidence-based variables affecting physicians' test-ordering tendencies: a systematic review. Neth J Med 2007;65:167–77.
15. Gross CP, Vogel EW, Dhond AJ, Marple CB, Edwards RA, Hauch O, et al. Factors influencing physicians' reported use of anticoagulation therapy in nonvalvular atrial fibrillation: a cross-sectional survey. Clin Ther 2003;25:1750–64.