Académie Universitaire
Wallonie-Europe

UNIVERSITÉ DE LIÈGE

FACULTÉ DE MÉDECINE

Contribution to the statistical evaluation of data obtained in External Quality Assessment programmes

Wim COUCKE

Dissertation présentée
en vue de l'obtention du grade de
Docteur en Sciences de la Santé publique

2012

UNIVERSITY OF LIEGE

Faculty of Medicine

Department of Public Health

Medical Informatics and Biostatistics

Contribution to the statistical evaluation of data obtained in External Quality Assessment programmes

Wim COUCKE

Supervisor
Professor A. Albert

January 2012

Acknowledgements

This work would never have been accomplished without the help, support and stimulation of many persons, to whom I want to express my deepest and honest gratitude.

Professor Adelin Albert, Head of Medical Informatics and Biostatistics, Faculty of Medicine, University of Liège, acted as my promotor and supervised my thesis. Prof. Albert, you have read this work several times in detail, commented and improved it. Your ideas and support were of priceless help, and your guidance with the redaction of the articles that have emerged from this work has been much appreciated. Also your continuous motivating words during the last years have kept me on track with the thesis.

I a much indebted to Professor Jean-Paul Chapelle, President of the accompanying committee and of the jury, as well as to the other members of the jury, for their interest in this work.

Nothing of this work would have been started without Dr. Jean-Claude Libeer, former head of the department of Clinical Biology of the Scientific Institute of Public Health (WIV-ISP). Dr. Libeer, you accepted me as a coworker in your department and you introduced me to the domain of External Quality Control. Your ideas and suggestions have been of great importance in this work.

I also want to thank Dr. Christel Van Campenhout and Dr. Philippe Van de Walle for contributing to my experience in External Quality Assessment programs and their sustained support at WIV-ISP.

My sincere gratitude also goes to Dr. Marjan Van Blerk. Marjan, you have introduced me to the art of writing articles, have let me participate in some of your research projects and let me have access to the surveys you were responsible for. Your ideas and practical assistance stimulated the development of several novel techniques.

I am much grateful to Dr. Nicole Hamers, who originally guided my ideas in external quality assessment and triggered the estradiol/progesterone article.

My cordial thanks also to Dr. Bernard China, who gave me new insights in the surveys he was conducting and offered me the opportunity to analyse the data.

I also want to express my gratitude to Professor Johan Smitz from the VUB Laboratory of Clinical Chemistry and Radioimmunology, who helped me for the estradiol/progesterone article.

All my colleagues of the department of Clinical Biology have to be thanked for their interest in the work and for creating the right atmosphere in which research can grow.

A special thanks goes also to Dr. Lixin Zhang for the interesting discussions we have had and for the sharing of ideas and new developments.

At last, Jesica, I am so grateful to you. You have been supporting me from the beginning of this work. You have shown patience and understanding, evening after evening, weekend after weekend of the past year and so many times before. Whenever I was hidden behind my computer screen, you completed with a smile what I had planned to do in the house.

Summary

Laboratory medicine has undergone a spectacular evolution in the last decades and has become today of crucial importance for supporting diagnostic and therapeutic decisions. The increase of the volume of laboratory analyses has not gone without an emerging risk of measurement errors that may have far-reaching consequences, even on the patient's life. External Quality Assessment (EQA), already established since several decades in various countries and often running on an international level, aim at going further than the "internal quality control" procedures of every laboratory and at improving laboratory quality by inter-laboratory comparisons. An EQA round generally consists of sending aliquots of the same sample to various laboratories for assaying selected tests. After finishing the assays, results are reported back to the EQA organizer. Subsequently these results are subject to a statistical analysis, which is performed globally, for all the participants, or for each analytical technique separately. Finally, a report is sent to every participant that informs about the acceptability of the individual results, with respect to predefined limits, and with respect to the group of peers.

This thesis, structured in five chapters, focuses on the External Quality Control of clinical laboratories by a critical analysis of existing methods and by

creating new approaches that permit to improve the current procedures.

The first chapter of this work emphasizes the evolution of the role of the clinical laboratory and EQA in the quality improvement. After the report 'To Err is Human: Building a Safer Health System', numerous scientists became interested in investigating the frequency, source and impact of laboratory errors. The Total Testing Process (TTP) became recognized as the best framework to investigate laboratory errors. The three different phases of the TTP - respectively, the pre-analytical, analytical and post-analytical phases - are described in detail and the nature and frequency of errors in each phase explained. For each phase, possible improvements are described and the role of EQA is suggested. Today, EQA principally focuses on the assessment and improvement of the analytical phase. Proposals are made to improve the role of EQA for assessing and improving pre- and post-analytical error as well, by using specific sample material and by automating the reporting of data and laboratory reports to the EQA participants. The principle of the comparison of results of a laboratory with those obtained by the other laboratories is traditionally based on the calculation of "z-scores". An in-depth study comparing different techniques has been made, shedding new light on the shortcomings and strong points of the different approaches. We concluded that robust techniques may exhibit weak performance for smaller sample size, while techniques that eliminate outliers before calculating z-scores should be recommended.

The second Chapter discusses the role of EQA as a tool to assess harmonization between methods. The role of EQA is described, together with the pitfalls and current shortcomings for assessing harmonization. A major problem in assessing standardization between methods is the possible presence of matrix effects in control samples, in which a method-specific bias may appear. Several explanations for matrix effects are mentioned and statistical techniques are described that assist EQA organizers to split up the data in homogeneous peer groups using multivariate statistics. The chapter also reviews several techniques to be used in method comparison studies, and the preference for the use of orthogonal regression is expressed. In addition, an example is given of a method-comparison study for Estradiol and Progesterone, with a novel technique of assessing standardization between various methods, in the presence of matrix effects for a small number of samples. The study also reveals that standardization between various methods is not attained, and that the striving for standardization with standards of higher order may not be satisfactory.

Chapter 3 introduces different evaluation techniques that combine information from different samples or parameters: Variance and bias index scores, Mean ranking scores, counts of z- and u-scores, and a long-term analytical Coefficient of Variation. Also, a new and original method is introduced that uses 3 steps to identify outliers in a first step, to find laboratories with exceeding variability in a second, and to identify laboratories with high bias in a third step. Each of the techniques are evaluated and discussed by means of a data set in which accidental outliers, high variability and high bias were induced. In addition, the comparison between the different evaluation methods reveals that distinguishing between variability and bias is a tedious task, and that some long-term analysis methods lack robustness against outliers. Also, it is proven that evaluation techniques summarizing results of different parameters may hide useful information. In addition, the 3-step method is proposed as a method for discerning between errors produced in the pre- or post-analytical phase, and errors that arise from the analytical phase.

Chapter 4 applies the 3-step method to data obtained from the Belgian EQA. Data sets from alcohol, flow cytometry, lithium and semen analysis surveys are examined. The method is extended for applicability to heteroscedastic, i.e. unequal residual variability, regression models and demonstrates that it is able to be used in a wide range of surveys. For each of the surveys under consideration, a follow-up is made of the occurrence of accidental mistakes, and the evolution of within-laboratory variability and bias for selected methods. It highlights several conclusions that show a striking similarity for various EQA surveys: an improvement of laboratory performance has been attained over time. The major improvement was a reduction of accidental mistakes. The analytical performance of selected methods, however, did not show an improvement over time.

In Chapter 5, some graphical representations of EQA data are explored and a graphical representation of the 3-step method is described. The histogram, normal quantile plot and box plot are described in detail and suggested for providing a quick visual overview of EQA data. Other graphical representations that respond to specific questions are given and discussed as well, like Shewhart charts, Cusum charts and graphical representations to combine variability and bias in one graph. In addition, the 3-step method is graphically explored by means of three distinct graphs. The chapter finishes by suggesting the use of interactive graphs for improving feedback from the EQA organizers to the EQA participants by means of Scalable Vector Graphics. The latter is illustrated with web-accessible examples of long-term evaluation of z-scores and the results of the 3-step method for the data obtained in the Belgian EQA for alcohol determination in blood.

In brief, this work describes in a critical and constructive way current statistical methods used in EQA and proposes novel statistical and graphical techniques to help alleviating the future needs of External Quality Assessment programmes.

Résumé

La biologie clinique a connu au cours des dernières décennies une évolution spectaculaire. Elle est devenue aujourd'hui une discipline incontournable dans l'aide à la décision médicale. L'augmentation du volume d'analyses de laboratoire n'est toutefois pas sans entraîner un risque accru d'erreurs de mesure pouvant avoir des conséquences graves sur la vie même des patients. Les programmes d'Evaluation Externe de la Qualité (EEQ), mis en place depuis plusieurs années tout au niveau national qu'international, se sont fixés pour objectifs d'aller au-delà des procédures de « contrôle de qualité interne » de chaque laboratoire et de contribuer à l'amélioration globale de la qualité par l'organisation d'enquêtes inter-laboratoires. Au cours de chaque enquête EEQ, un (ou plusieurs) même échantillon contrôle est envoyé à l'ensemble des laboratoires pour analyse. Une fois les dosages terminés, les résultats sont renvoyés à l'organisme responsable EEQ. Ces résultats font ensuite l'objet d'une analyse statistique, globale ou par technique de dosage (c'est-à-dire par groupe de laboratoires utilisant le même principe analytique) et un rapport est adressé à chaque participant l'informant sur l'acceptabilité de ses résultats (par rapport à des limites prédéfinies) et sur sa performance par rapport à ses pairs.

Ce mémoire, structuré en cinq chapitres, s'est intéressé à l'Evaluation Externe de la Qualité des laboratoires de biologie clinique par une analyse critique des méthodes existantes ainsi que par la mise au point de nouvelles approches permettant d'améliorer les procédures actuellement en place.

Au Chapitre 1, l'accent est mis sur l'évolution du rôle des laboratoires de biologie clinique et des programmes EEQ dans l'amélioration de la qualité. Après la publication du rapport « To err is human: building a safer health system », de nombreux chercheurs ont commencé à s'intéresser et étudier en détail la fréquence, l'origine et l'impact des erreurs de laboratoire. Dans ce contexte, le concept de « Total Testing Process (TTP) » est reconnu comme le meilleur système pour traquer les erreurs de laboratoire. Les trois phases du TTP, respectivement libellées pré-analytique, analytique et post-analytique, sont détaillées en spécifiant pour chacune d'elles la nature et la fréquence des erreurs. De plus, des améliorations potentielles sont décrites à chaque fois, notamment grâce au rôle que peuvent jouer les programmes EEQ dans cette problématique. Il faut reconnaître qu'à l'heure actuelle, ces derniers se focalisent surtout sur l'évaluation et l'amélioration de la phase analytique. Quelques propositions sont faites pour améliorer le rôle des programmes EEQ dans le suivi et le développement des phases pré-analytique et post-analytique, en utilisant des échantillons contrôles de matériel spécifique et en automatisant l'échange des données et des rapports entre les participants et l'organisme EEQ. Le principe de la comparaison des résultats d'un laboratoire avec ceux obtenus par les autres laboratoires est traditionnellement basé sur le calcul des « z-scores ». Une étude comparative des méthodes de calcul des z-scores a été menée afin de mettre en évidence les points forts et les points faibles de chacune d'elles. Il en ressort que les méthodes basées sur la statistique robuste sont peu fiables lorsque le nombre de laboratoires est faible, tandis que celles basées sur l'élimination préalable des valeurs aberrantes (« outliers ») avant le calcul des z-scores sont meilleures et dès lors recommandées.

Le Chapitre 2 s'intéresse à l'intérêt des programmes EEQ en tant qu'outil d'évaluation de l'harmonisation des méthodes analytiques. Le rôle de l'EEQ, les pièges et limitations actuelles de l'évaluation de l'harmonisation y sont décrits. On évoque notamment la problématique de l'effet de matrice dans les échantillons contrôles, qui peuvent causer un biais spécifique par méthode. Ces effets de matrice trouvent des explications et peuvent être contournés en formant des groupes homogènes de laboratoires en ayant recours à des techniques de la statistique multivariée. Le chapitre 2 passe aussi en revue plusieurs approches statistiques pour comparer différentes méthodes analytiques; une préférence se dégage pour l'utilisation de la régression orthogonale. En guise d'illustration, une étude comparative des méthodes de dosage de l'estradiol et de la progestérone a été menée pour essayer de standardiser plusieurs méthodes de dosage en présence d'effets de matrice pour certains échantillons. L'étude montre que la standardisation n'est pas atteinte et que le fait de rechercher une standardisation en se basant sur es standards d'ordre plus élevé pourrait s'avérer insatisfaisant.

Le Chapitre 3 introduit diverses techniques d'évaluation de la qualité combinant l'information de plusieurs échantillons ou paramètres, à savoir les scores de variabilité et de biais, le score basé sur le classement des moyennes, les scores Z et U, et le coefficient de variation (CV) à long-terme. On propose également une méthode nouvelle et originale en trois étapes (« 3-step method ») basée sur le modèle de régression linéaire. La première étape consiste en la détection d'éventuels « outliers» pour chaque laboratoire ; dans une deuxième étape, on identifie ensuite les laboratoires ayant une variabilité excessive ; enfin dans une troisième étape, on repère les laboratoires avec un biais important. Les méthodes existantes précitées et la nouvelle approche en 3 étapes font ensuite l'objet d'une étude comparative par simulations sur un échantillon de données dans lequel on a introduit des erreurs accidentelles, une variabilité excessive et un biais élevé. Les simulations montrent qu'il n'est pas facile de distinguer entre la variabilité et le biais, et que certaines méthodes d'analyse à long-terme ne sont pas robustes en présence d'« outliers ». On a aussi montré que les techniques d'évaluation des différents paramètres peuvent masquer des informations utiles. Le point saillant de l'étude est que la méthode en 3 étapes permet de distinguer les erreurs qui surviennent lors des phases pré-analytique et post-analytique de celles provenant de la phase analytique.

Au Chapitre 4, la méthode en 3 étapes est appliquée à des données du programme Belge de l'Evaluation Externe de la Qualité : alcoolémie, cytométrie de flux, lithium et analyse de sperme. La méthode est ensuite étendue pour être applicable à des situations où les conditions d'homoscédasticité (homogénéité des variances) ne sont pas respectées, situation fréquente en EEQ. On a effectué un suivi de l'occurrence d'erreurs accidentelles et de la variabilité et du biais inter-laboratoires pour une méthode donnée des programmes EEQ belges. Les conclusions montrent une similarité remarquable entre les différentes périodes EEQ, à savoir une amélioration globale des prestations des laboratoires dans le temps, due essentiellement à la réduction des erreurs accidentelles. En effet, la performance analytique des méthodes de dosage n'a pas réellement montré d'amélioration.

Le Chapitre 5 est consacré aux représentations graphiques des données EEQ, en particulier pour la méthode en 3 étapes. L'histogramme, le graphe des quantiles normaux et la boîte à moustaches sont fréquemment utilisés pour une exploration visuelle des données EEQ. D'autres représentations graphiques, davantage liées à des questions spécifiques, sont mentionnées comme la carte de contrôle de Shewhart, celle des « cusums » et les représentations qui combinent la variabilité et le biais sur un seul graphique. La méthode en 3 étapes est explorée graphiquement par trois graphes séparés. Le chapitre se termine avec la présentation de graphes interactifs comme moyen d'améliorer le feedback des organisateurs EEQ aux participants, notamment en ayant recours aux graphes vectoriels adaptables. Ceci est illustré par des exemples qui peuvent être visualisés sur le web, notamment une évaluation à long-terme des z-scores et les résultats de la méthode en 3 étapes pour les données EEQ d'alcoolémie dans le sang.

En conclusion, ce travail revisite de façon critique et constructive les méth-

odes statistiques et graphiques utilisées actuellement en EEQ et propose de nouvelles approches visant à répondre aux besoins futurs des programmes d'Evaluation Externe de la Qualité.

Samenvatting

Laboratoriumgeneeskunde heeft de laatste decennia een spectaculaire evolutie ondergaan en is tegenwoordig een onmisbare discipline geworden voor de ondersteuning van diagnostische en therapeutische beslissingen. De forse stijging van het aantal laboratoriumanalyses is samengegaan met een verhoogd risico op meetfouten, die verreikende gevolgen, ook voor het leven van de patiënten, kunnen hebben. Externe kwaliteitscontrole (EKE), een discipline die al gedurende verschillende jaren in verschillende landen en op een internationaal niveau wordt uitgevoerd, is één van de mechanismes om de kwaliteit in het laboratorium te verbeteren, om een aanvulling te zijn op de interne kwaliteitscontrole van elk laboratorium en om bij te dragen aan de algemene verbetering van de kwaliteit door de organisatie van interlaboratoriumproeven. Een EKE-ronde bestaat in het algemeen uit het versturen van de aliquots van hetzelfde staal naar verschillende laboratoria, die bepaalde testen erop uitvoeren. Na de analyse worden de resultaten teruggestuurd naar de EKE-organisator. Daarna worden de resultaten statistisch geanalyseerd, globaal of per meettechniek (dwz per groep van laboratoria die dezelfde meettechnologie gebruiken) en een rapport wordt verstuurd naar elke deelnemer dat informeert over de aanvaardbaarheid van zijn resultaten, tegenover vooraf gestelde limieten en tegenover de andere laboratoria die eenzelfde

meettechniek gebruiken.

Dit werk, in vijf hoofdstukken opgevat, gaat dieper in op de Externe Kwaliteitsevaluatie van de klinische laboratoria door een kritische analyse van de bestaande evaluatiemethodes en ook door het opstellen van nieuwe benaderingen die in staat zijn om de huidige evaluatietechnieken te verbeteren.

In het eerste hoofdstuk ligt de klemtoon op de evolutie van de rol van het klinisch laboratorium en EKE. Nadat het IOM rapport 'To Err is Human: Building a Safer Health System' was verschenen, zijn wetenschappers begonnen de frequentie, bron en impact van laboratoriumfouten in detail te onderzoeken. Het totale testproces (TTP) wordt beschreven als het best referentiekader om laboratoriumfouten te onderzoeken. De drie verschillende fases van het TTP - respectievelijk pre-analytisch, analytisch en post-analytisch - worden in detail beschreven en het karakter en de frequentie van fouten in elk proces wordt uitgelegd. Voor iedere fase worden toekomstige verbeteringen beschreven en de rol van EKE wordt hierin voorgesteld. Tegenwoordig bekommert de EKE zich vooral om de evaluatie en de verbetering van de analytische fase. Er worden enkele voorstellen gemaakt waarbij EKE programma's ook het pre- en post-analytisch deel kunnen opvolgen, door specifiek staalmateriaal te gebruiken en door het rapporteringssyteem voor gegevens van de laboratoria naar de EKE organisator te automatiseren. Het principe van de vergelijking van resultaten van een laboratorium met deze bekomen door de andere laboratoria is traditioneel gebaseerd op de berekening van "z-scores". Een studie die verschillende berekeningswijzes van z-scores vergelijkt is uitgevoerd teneinde de sterke en zwakke punten van elke wijze bloot te leggen. Deze studie leidde tot de conclusie dat de methodes die op robuuste statistieken gebaseerd zijn, minder betrouwbaar zijn wanneer het aantal laboratoria laag is, terwijl zij die gebaseerd zijn op een voorafgaande stap die afwijkende waardes ("outliers") elimineert betrouwbaardere z-scores opleveren en dus aangeraden zijn.

Het tweede hoofdstuk beschrijft de rol van EKE als een middel om harmonisering tussen methodes na te gaan. De rol van EKE wordt beschreven, samen met de valkuilen en huidige tekortkomingen om harmonisering na te gaan. Een groot probleem hierbij is het verschijnsel van matrixeffecten in controlestalen, die een methode-specifieke bias kunnen veroorzaken. Verschillende verklaringen voor matrixeffecten worden gegeven, en er wordt een beschrijving gegeven van statistische technieken die EKE-organisatoren informatie kunnen verschaffen of de te evalueren groepen homogeen zijn. Oplossingen worden voorgesteld om de te evalueren groepen op te splitsen aan de hand van multivariate statistiek. Het tweede hoofdstuk beschrijft ook verschillende technieken die kunnen gebruikt worden in studies die methodes met elkaar vergelijken, en een voorkeur voor orthogonale regressie wordt uitgelegd. Daarnaast wordt een voorbeeld gegeven van een studie die methodes voor Estradiol en Progesterone vergelijkt met een nieuwe techniek om standaardisering na te gaan tussen methodes in aanwezigheid van matrixeffecten bij een klein deel van de stalen. De studie toont ook aan dat standaardisering tussen verschillende methodes niet bereikt is, en het streven naar standaardisatie op basis van standaarden van hogere orde onvoldoende zou kunnen zijn.

Het derde hoodstuk introduceert verschillende evaluatietechnieken die informatie combineren van verschillende stalen of parameters: variantie en bias index scores, scores gebaseerd op de gemiddelde rangorde, tellingen van z- en u-scores, en een langetermijns- analytische variatiecoëfficiënt. Er wordt ook een nieuwe en originele methode voorgesteld die is gebaseerd op het lineaire regressiemodel gebruikt en in 3 stappen ("3-step method") werkt. De eerste stap spoort eventuele outliers op voor elk laboratorium apart, in een tweede stap worden de laboratoria geïdentificeerd met een te grote variabiliteit, en in een derde stap worden de laboratoria geïdentificeerd met een grote bias. Dit model wordt in detail beschreven en zijn voor- en nadelen worden gedetailleerd besproken. Elke zonet vermelde techniek wordt geëvalueerd en besproken aan de hand van een dataset waarin toevallige fouten, hoge variabiliteit en extreme bias werden gestoken. De vergelijking tussen deze methodes toont aan dat het niet eenvoudig is om een onderscheid te maken tussen variabiliteit en bias, en dat sommige methodes voor een analyse op lange termijn niet robuust zijn tegen de aanwezigheid van "outliers". Er wordt ook aangetoond dat evaluatietechnieken die de resultaten van verschillende parameters omvatten soms nuttige informatie verbergen. Ten slotte wordt de 3-stapsmethode voorgesteld als een methode om een onderscheid te maken tussen de fouten die in de pre- en post-analytische fase worden gemaakt, en fouten die ontstaan bij de analytische fase.

In het vierde Hoofdstuk wordt de 3-stapsmethode toegepast op gegevens die bekomen werden in het Belgische systeem voor Externe Kwaliteitsevaluatie voor alcohol, flow cytometrie, lithium en sperma-analyse. De methode wordt uitgebreid om toegepast te worden voor heteroscedastische regressiemodellen en er kan gezien worden dat ze toepasbaar is in een breed domein van EKE's. Voor elk van de EKE's die in beschouwing worden genomen wordt een opvolgstudie gemaakt van het voorkomen van toevallige fouten, en de evolutie van de variabiliteit tussen de laboratoria en de bias voor enkele methodes. De conclusies vertonen een opvallende gelijkenis voor verschillende EKE's: een verbetering van prestaties van laboratoria is bereikt over de tijd. De voornaamste verbetering was een reductie van het aantal toevallige fouten. De prestatie van de analytische methodes vertoonde echter geen verbetering over de tijd heen.

Het vijfde Hoofdstuk wijdt zich aan de grafische voorstelling van EKE-gegevens en een beschrijving van een grafische voorstelling van de 3-staps methode. Het histogram, de normale quantielplot en de box plot worden vaak gebruikt voor een visuele verkenning van EKE gegevens. Andere grafische voorstellingen, die eerder te maken hebben met specifieke vraagstellingen, worden ook gegeven en besproken, zoals Shewhart grafieken, Cusum grafieken en grafische voorstellingen die variabiliteit en bias in één grafiek combineren. Daarnaast wordt de 3-staps methode grafisch verkend door middel van drie aparte grafieken. Het hoofdstuk eindigt met de voorstelling van interactieve grafieken als een middel om de terugkoppeling te verbeteren van EKE organisatoren naar de deelnemers door middel van Schaalbare Vectorgrafieken. Dit wordt geïllustreerd met voorbeelden van een lange-termijn evaluatie van z-scores en de resultaten van de 3-stapsmethode voor de gegevens bekomen in het Belgische EKE voor bepaling van alcohol in bloed en die via het web kunnen worden bekeken.

Kortom, dit werk beschrijft op een kritische en constructieve manier de verschillende statistische en grafische technieken die momenteel in EKE worden gebruikt en beschrijft nieuwe benaderingen die een antwoord kunnen geven op de toekomstige noden van programma's voor Externe Kwaliteitsevaluatie.

Contents

Glossary

ACD	Analytical Critical Difference
AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
B	Bias
BIS	Bias Index Score
CB	Constant Bias
CCV	Chosen Coefficient of Variation
CLSI	Clinical and Laboratory Standards Institute
CV	Coefficient of variation
DOM	Document Object Model

DV	Designated Value
E_2	Estradiol
ECAT	European Concerted Action on Thrombosis
ECDF	Empirical Cumulative Distribution Function
EDTA	Ethylenediaminetetraacetic acid
EQA	External Quality Assessment
g/L	gram per liter
GLM	Generalized Linear Model
GLMM	General Linear Mixed Model
GUM	Guide to the expression of Uncertainty in Measurement
HAMA	Human Anti-Mouse Antibodies
HL7	Health Level Seven International
HTML	Hyptertext Markup Language
ICP-MS	Inductively Coupled Plasma-Mass Spectrometry
ILAC	International Laboratory Accreditation Cooperation
IOM	Institute of Medicine
ISO	International Organization for Standardization
ID-GC/MS	Isotope Dilution - Gas Chromatography/Mass spectrometry

IVD	In Vitro Diagnostic Medical Devices
IQR	Interquartile range
JVM	Java Virtual Machine
$\mathrm{LCV_a}$	Long-Term Analytical Coefficient of Variation
Li	Lithium
LIS	Laboratory Information system
LOINC	Logical Observation Identifier Names and Codes
LTS	Least Trimmed Squares (regression)
MANOVA	Multivariate Analysis Of Variance
MCD	Minimum Covariance Determinant
mL	milliliter
mmol/L	millimol per liter
MRBIS	Mean Running Bias Index Score
MRVIS	Mean Running Variance Index Score
nmol/L	nanomol per liter
NPV	Negative Predictive Value
OLS	Ordinary Least Squares (regression)
OMRVIS	Overall Mean Running Variance Index score

P	Progesterone
PB	Pillai-Bartlett trace
PDF	Portable Document Format
pmol/L	picomol per liter
PPV	Positive Predictive Value
PrB	Proportional Bias
PSA	Prostate-Specific Antigen
RM	Reference Material
SDs	Standard Deviation(s)
SI	Système International d'unités (International System of Units)
SSE	Residual Sums of Squares
SVG	Scalable Vector Graphics
SWF	Shockwave Flash
TE	Total Error
TTP	Total Testing Process
U/L	Units per liter
VIS	Variance Index score

W3C	World Web Consortium
WIV-ISP	Wetenschappelijk Instituut Volksgezondheid - Institut Scientifique de Santé Publique
XML	Extensible Markup Language

CHAPTER 1

External Quality Assessment for Laboratory medicine

1.1 Introduction

Laboratory medicine, the specialty that deals with assaying biological specimens from patients, is an integral part of the complex process of therapy control and management of a patient's disease [137, 146]. It is estimated that laboratory test results have an impact on over 70 percent of medical decisions [54, 15]. Clinical laboratories are high-output, high-quality entities, employing highly skilled and continuously trained staff [146, 46, 83]. They do not only deliver test results, they also act as knowledge service and translate laboratory data into comprehensive information for the clinician [84, 20]. During the last decade, clinical laboratories have undergone fundamental changes [16]. Their complexity and role in health care has evolved, requiring a greater involvement of laboratory professionals in the diagnostic process [46]. Remarkable developments in instrument technology and computer science, laboratory automation, the expansion and improvement of testing procedures and controls, and compliance with systems of quality management, have greatly enhanced the possibilities of the clinical laboratory. Nonetheless, technological advances in laboratory medicine have also taken place in a context of drastic cost reductions. Two types of approaches have been identified to reduce costs. The first approach focuses on advantages of scale and encompasses consolidation of laboratory sections with the creation of big central core laboratories, largely relying on automation and scale savings. Secondly, an increased efficiency has been reached by improving the diagnostic performance, creating more effective diagnostic strategies and effectively utilizing laboratory information for the diagnosis and treatment of patients [119, 112, 145].

In spite of the changing needs and technologies, the basic process of the clinical laboratory has remained the same and it can be precisely described by the Total Testing Process (TTP), consisting of 11 steps divided into three phases, first devoloped by Lundberg [88]. The pre-analytical phase, ranging from taking the sample to preparing it for analysis; the analytical phase, during which a test result is produced, and finally the post-analytical phase, in which the result is verified, interpreted and a medical action is taken (see Fig 1.1). Recently, it has been recognized that the pre-analytical phase is preceded by a pre-pre-analytical phase and the post-analytical phase is succeeded by a post-post-analytical phase. The “pre-pre-analytical” phase includes the formulation of a clinical question and the selection of appropriate laboratory tests. In the “post-post-analytical” phase, the clinician interprets laboratory results to build a diagnosis and take subsequent therapeutic steps [163].

Traditionally, clinical laboratories have been concerned with the quality of the analytical aspect of their work. Unlike many other medical processes, these activities are precisely defined and more controllable than a procedure or treatment in other medical disciplines. In recent decades, standardization, automation and technological advances have significantly improved the analytical reliability of laboratory results and decreased the error rates to a level

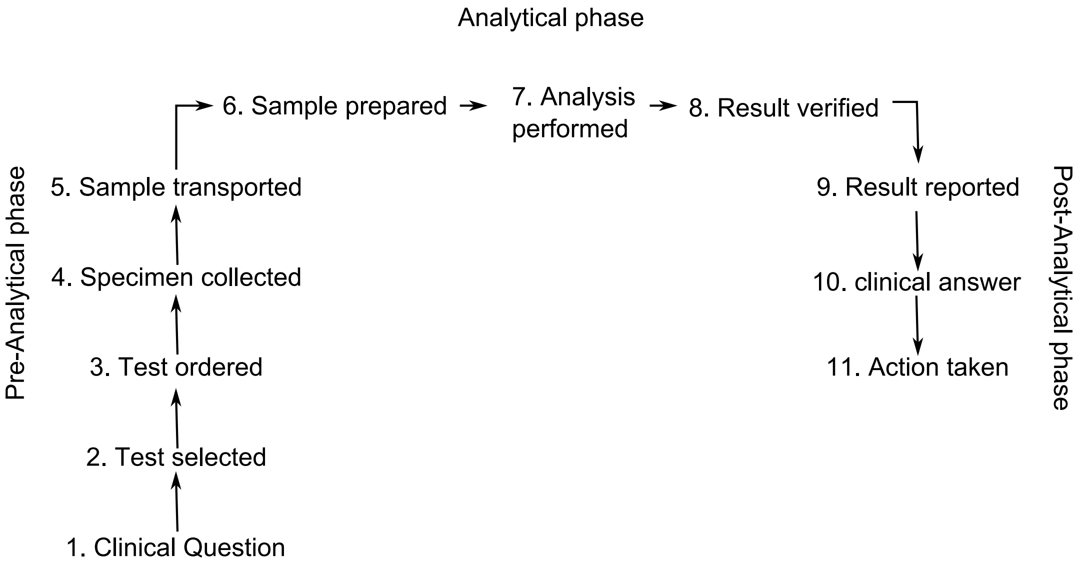

Figure 1.1 The 11-step process of a clinical laboratory analysis.

which is far lower than seen in overall clinical health care areas, although it doesn't match up yet with industrial quality standards [185, 68, 123, 106].

In 1999, the Institute of Medicine (IOM) published its report 'To Err is Human: Building a Safer Health System' [71], investigating the level of mistakes in US clinical laboratories. It indicated that medical errors were the eighth most important cause of death in the USA, more important than breast cancer or AIDS-related deaths. It was the start of a debate and concern about patient injuries in health care. Patient safety, an issue not well understood or discussed in health care systems, became a major subject of discussion and clinical laboratories started to investigate the nature, causes and frequency of their errors. In this chapter, the role of External Quality Assessment in improving quality and reducing laboratory errors will be discussed, with highlights on the current role of External Quality Assessment (EQA) in the changing environment of laboratory medicine. The Total Testing Process will be used as a framework for exploring sources and solutions for mistakes. For each phase, the role of EQA will be exemplified and the chapter ends with an in depth discussion of z-scores, which are one of the most common statistical techniques to evaluate EQA data.

1.2 Errors in laboratory medicine

1.2.1 Introduction

Investigating errors in laboratory medicine should start from a patient-centered approach. Any direct or indirect negative consequence for the patient related to a laboratory test must be considered [125, 127]. The Total Testing Process is an excellent framework to start from. A mistake can occur in each of the 11 steps in this process [15], starting from test request and ending with the physician's reaction to laboratory information. According to this perspective, the definition of laboratory error is ‘‘a defect occurring at any part of the laboratory cycle, from ordering tests to reporting results and appropriately interpreting and reacting on these". Investigating errors is nevertheless a complicated subject, since the Total Testing Process is complex, consisting of a series of interrelated processes [122, 13, 79, 128].

1.2.2 Identification of laboratory errors

Pre-analytical phase

The pre-analytical phase includes the phases from the test request to the preparation of the sample for analysis [122]. Specifically, it starts from the clinician's request, including the selection of tests to be performed, preparation of the patient, collection of the primary sample, continuous with the transportation of the sample to and within the laboratory and ends just before the analytical procedures begin. By far, sample misidentification is a serious concern [81]. In addition, several patient-related and less controllable variables may influence the in vivo concentration of several parameters in the patient's body fluids, for example the patient's age, gender, ethnic background, diet, physical exercise or drug use. They need to be considered when interpreting the final laboratory result and hence they have to be registered correctly before sampling [25, 146, 80, 124]. The act of taking the sample may also influence the laboratory results. Posture during blood sampling for example, influences numerous parameters, and the use of a tourniquet may change the concentrations of various analytes. The transportation of the sample from the sampling site to the laboratory has to ensure a timely delivery as well [146]. The type of container the blood is collected is a frequent mistake as well [85]. In case of pre-analytical mistakes, the concentration of the analyte in the sample does not reflect the true concentration in the patient's blood or other body fluids, although the sample may have been analyzed correctly [146].

Analytical phase

The analytical phase starts when a sample is ready for analysis and lasts till the approval of the test result. It consists of the sample preparation, the determination of the test result, mostly by an automated analyzer, and the final approval by the medical laboratory staff. The performance of analytical procedures is often described in terms of bias and variability, or trueness and precision. Trueness is the proximity between the average value obtained from a large series of test results and an accepted reference value [97]. For method with high trueness, the average of several measurements made on the same sample will be very close to the true value. For a precise method, repeated measurements will be very close to each other. It should be noted that, although different factors may influence trueness and precision in a different way, they both need to be under control to ensure the delivery of a high quality test result.

Problems may arise before analysis when samples have to be diluted, for example when the sample volume is insufficient or the dilution is inappropriate. Almost every analysis is preceded by a calibration and an internal control check. Calibrators may be out of date or prepared in a wrong way before analysis. One way to control the analytical process is internal quality control. Adequate action should be undertaken whenever the internal quality control indicates that the analysis process is out of specifications. In addition, every analytical method should give comparable results to any other analytical method and in particular to a reference method. Although reference methods with well-defined standards are available for many methods, a problem persists for different antibody based assays. In this context, it is important to use appropriate statistical techniques when performing method comparisons. Often, manufacturers rely on Pearson correlation coefficients and interpret them in conditions when they should not be used. Moreover, interference with analytical procedures by endogenous or exogenous substances may lead to false results [156]. Hemolysis, for example, is one of the factors that may impact the final test results. Although debatable, we prefer to classify hemolysis as an analytical event, since it is not always possible to recognize it before analysis and because it interferes with the measurement of several analytes [166]. The most important example from exogenous substances which may interfere with analytical test results are drugs and their metabolites, and substances added to the blood to avoid coagulation [146].

Postanalytical phase

The postanalytical phase starts when the result has been verified and lasts till the appropriate action has been taken by the medical staff. It includes verification of laboratory results by qualified personnel, storing them into the laboratory information system and reporting them to the clinical staff who requested the tests. Often interpretative comments are given to guide the clinical staff in their final diagnostic decision. It also involves an alert system by passing urgent messages to the clinical staff whenever necessary. Reporting of results strongly affects the effective translation of results into clinical information [70]. It also includes storage and disposal of samples when the analyses have been completed [146, 120]. However, most post-analytical errors derive from inappropriate interpretation and utilization of laboratory results. A part of the post analytical procedures take place outside of the laboratory and may be described by post-post analytical steps. They include the receival, reading and interpretation of the results by the clinician, and the final decision of the clinician based on the information provided to him.

1.2.3 Frequency of laboratory mistakes

During the last decades, the laboratory error rate has been significantly reduced, in particular for analytical errors. The variability of analytical performance is now less than 1/20th of what it was 40 years ago. Further, it has been estimated that up to 75% of laboratory errors generate results without endangering the patient results, and that 10-20 % are absurd results which are detected before they result in inappropriate care. There is a rest of less than 10% of errors which may have an impact on patient care [55, 126, 128]. Looking at the different phases of the TTP, it is generally accepted that the pre-analytical phase is the most error-prone, with reported frequencies ranging from 60 to more than 80% [126, 122, 55, 166, 21, 82]. Even more, the frequency of error in this phase, which still consists of manually intensive procedures, has increased in the last decade [79]. The error frequency of the analytical phase is around 10%, and ranges in the postanalytical phase between 20 and 50 %. It should be noted however that the exact number of mistakes is difficult to assess, because there exists no widespread process to follow them up systematically and there is a wide variety of definitions and methods to measure them. Moreover, it has been shown that errors are reported with higher frequency when studies look particularly for them than when they are reported in the daily routine. Also, there may be a biased focus on mistakes that result in adverse events [15]. Overall, the reported error rate in clinical laboratories ranges from less than 0.05% up to 10% [68, 21].

1.2.4 Reducing errors

The Institute of Medicine report “To Err Is Human” generated widespread interest in medical errors and adverse events in health care, as well as strategies for reducing them. The most obvious improvement in the latest decades may have been the widespread implementation of quality management systems, encompassing processes for process and risk analysis, and focussing on error prevention, detection and management [79, 124, 83]. Quality management systems should cover all the steps involved in the overall testing and non-testing processes [133] and the implementation of a systematic error tracking system may be worked out in this framework. The International Organization for Standardization (ISO) has released a norm, ISO 15189 [62], focusing on quality management in the clinical laboratory. It encompasses all the steps and processes within the TTP [120, 124, 21]. The following paragraphs will explore the three main phases of the Total Testing Process and describe in detail the different opportunities to improve quality in the clinical laboratory which are not directly a consequence of a quality system implementation.

Pre-analytical phase

Several issues have become important, such as involving the provision of appropriate clinical history and laboratory specimens, and the satisfaction of patients and professional staff about the collection process have become important [55]. Automation has been described as one of the major recent improvements to avoid errors in this phase. Bar-coded wrist bands, for example, reduce patient identification errors, pre-analytical workstations performing sample preparation tasks are seen as a major improvement, and software exists to reduce the number of mistakes during phlebotomy [128, 123, 85]. In addition, data loggers that register environmental parameters and three-dimensional acceleration may inform about unwanted acceleration of automatic transportation chain [37]. A recent study has shown, for example, that sudden acceleration had a direct relation with increased hemolysis [162].

Analytical phase

The analytical phase has been for decades the phase that has attracted most attention for improvement. In 1950, Levey and Jennings introduced the role of control charts in clinical chemistry and clinical laboratories [76]. Later on, Westgard et al developed a series of QC rules based on statistical process control to effectively apply internal control techniques [184, 183]. In addition to good test performance and result calculation, attention is now also drawn to adequate turnaround times and the electronic delivery of test result reports. Engineering developments have helped to create automated analyzers, which are able to perform routine analyzes in a more standardized way [166]. Traditionally, External Quality Assessment (EQA) schemes have been designed to control for errors in the analytical phase, providing participants a view of their analytical performance with respect to other laboratories and providing an overview of analytical performance of different types of analyzers.

Post-analytical phase

It has become essential for laboratory medicine to set high standards on the way they communicate with clinicians, since laboratory tests serve as a basis for many clinical decisions. Next to the test result, interpretation of results should be given. Also, any information which may explain a deviating result should be noted, like for a example a bias due to a high lipid or protein concentration in the sample [156].

Information technology, in particular expert systems, may be applied to screen and verify test results. They provide for example the means for a delta check, which compares the current test result with previous results from the patient. If the values are significantly different from historical values, the result is flagged. It is important here to note that the choice of parameters to perform delta checks should be made carefully and taking into account the type of patient population [161]. In addition, electronic communication and reporting offer significant opportunities to enhance the communication between the laboratory and clinicians in the field and reduce diagnostic errors [81, 55, 122, 123]. Care still has to be taken with keyboard entry errors, and, as always, processes supported by advanced automatizing rely heavily on a stable network infrastructure. The more processes rely on computerization, the more vulnerable they are to network downtime or computer failures. Attention should also be given to the interpretation of results by the clinicians. For example, a study [158] has shown that, even when specificity and sensitivity of test are communicated, clinicians do not always know how to interpret these values. Especially when the prevalence of a disease is low, diagnostic decisions may be impacted. To provide high quality interpretative comments, laboratory personnel should receive adequate training and also here the need for quality assurance and audit is high [14, 123].

1.3 Quality for the clinical laboratory

Quality assurance has been defined as the whole spectrum of quality improving activities which ensure the usefulness of laboratory investigation. The International Laboratory Accreditation Cooperation (ILAC) and the International Organization for Standardization (ISO) have released standards relating to this subject. The gold standard for most non-medical laboratories has been ISO 17025 [61]. The International Organization for Standardization has released a document particularly focusing on quality assurance in medical laboratories, ISO 15189 [62]. It describes among others the organization and management of the clinical laboratories, the set up of a quality management system and management of nonconformities. It focuses on the three phases of the Total Testing Process and describes in detail the actions to be undertaken to maintain and assure quality. It also clearly requires that laboratories participate in External Quality Assessment programmes [108, 153]. By participating, a laboratory is able to to check whether its performance is within the limits of its internal standards and to derive measures for continuously improving the quality of its work [18]. Results from External Quality Assessment may be regarded as the single most relevant indicator of laboratory quality [121].

1.4 External quality assessment

1.4.1 Background

External Quality Assessment is a widely used tool for assessing whether laboratories perform tests competently. Its objective is to ascertain and assess the ability of individuals and protocols to perform clinical assays satisfactorily, beyond the particular items or challenges presented. It may also monitor the quality of commercial analytical systems, reagents and test kits

[148, 169, 150]. It has its origin in the work of Belk and Sunderman [165], who distributed samples from the same origin to a group of laboratories in the Philadelphia area in 1947. They were originally used to identify a small number of incompetent practitioners and may have been punitive towards poor performers [108]. Nowadays External Quality Assessment is primarily educational, providing valuable information to individual participants relating to their performance as well as facilitating method comparisons. Satisfactory results in EQA schemes are an important evidence that analytical procedures are under control, that technicians work in an appropriate way, and that effective internal quality rules are in place [29, 43, 63, 154, 121, 42]. Two terms exist to describe this kind of activities, often used interchangeably: “proficiency testing” and “External Quality Assessment”. In general, the former is more often used in North America and often related with regulatory, or legal, attributes. The latter is more often used within the European area and is often seen as a broader activity, of which proficiency testing makes up only a part, and usually regarded as educational and as a self-assessment tool. In this work, we will stick to the term “External Quality Assessment”. Currently, External Quality Assessment schemes operate in the field of laboratory medicine in many countries and their aims, stages of development, and design differ [169]. EQA programmes have shown in the past their usefulness by helping to improve analytical performance and hence diagnostic accuracy. The educational aspect of EQA programmes has been mentioned several times as one of the key aspects helping to improve the quality of clinical laboratories. This is likely due to a combination of direct participant feedback and education through scientific meetings and publications [53, 69, 132]. In addition, the follow-up of poor performance in EQA programmes has helped laboratories in identifying analytical errors and detecting their possible sources [14]. Also, the elimination of poor performers has resulted in an overall improvement of test quality [69]. There is however only indirect evidence that EQA has improved patient safety [108].

Since EQA programmes are the main tool enabling laboratories to monitor the overall quality of their results, they must be aware of their own challenges and pitfalls. In particular, setting target values, preferably with traceability towards reference methods, commutability of patient and EQA samples, and providing comprehensive evaluation of the individual laboratory are cornerstones of a well functioning EQA programme. Also, they should promote the improvement of overall performances and support the management of unsatisfactory performances [149, 121]. For this reason, it is important that EQA organizers strive to use sample material which resembles real, single-donor samples as much as possible, while achieving a stability which is strong enough to cover the time delay by sending out the samples. Also, generating comprehensive information for laboratories, setting performance goals and providing information about trends in analytical techniques and performances are a continuous challenge [121].

A central and important aspect is the evaluation of the laboratories' outcome of measurements. Two main approaches exist to establish the performance of laboratories. The first approach is the so-called peer group comparison of each individual result with the results obtained by the other laboratories. It assumes that the majority of the laboratories answer a result that is equal or close to the correct result, and identifies laboratories who have responded deviating values as poor performers. The procedure to determine when a result deviates and to identify these laboratories will be explained in larger detail later on in this work. The second approach is based on analytical quality specifications or goals [67] which have been set by external criteria. Setting analytical quality specifications has been a subject of debate for many years and several strategies have been developed to calculate them. In 1999, a conference sponsored by IUPAC, IFCC and WHO was held in Stockholm on ''Strategies to Set Global Analytical Quality Specifications in Laboratory Medicine''. It was a milestone in the discussion on setting analytical target. The hierarchy of strategies can be summarized as follows [117]:

1. Evaluation of the effect of analytical performance on clinical outcomes in specific clinical situations. Sometimes a clear and unequivocal relation exists between levels of a certain single parameter and the onset of a particular

pathology, for example in markers for cardial damage or in lead poisoning. Performance limits may be derived from threshold values in these specific settings.

2. Evaluation of the effect of analytical performance on clinical decisions in general. Here, data based on components of biological variation or based on the analysis of clinician's opinion may be used.

3. Published professional recommendations from national and international expert bodies or from expert local groups or individuals. For several parameters, like serum cholesterol for cardial failure risk assessment, or for serum prostate-specific antigen (PSA), cut off values are determined by an international expert group and may be used as a basis for calculating specification limits.

4. Performance goals set by regulatory bodies or organizers of External Quality Assessment schemes. The US national External Quality Assessment organizer CLIA has established its own limits to evaluate its participants. The procedures to develop these limits are not explicitly provided, they are implicitly linked to the state-of-the-art laboratory practice as of 1988.

5. Goals based on the current state of the art, demonstrated by data from EQA or as found in current publications about the applied methodology. EQA organizers may use the data of their surveys to calculate limits beyond which a normal value is very unlikely to fall. These limits however are purely empirical and don't have any link to clinical decision-making or biological variability.

In 2010, the main actors in this consensus strategy have evaluated this procedure and concluded that the strategies developed then have shown to be successful. They also recognized that work was still to be done, mainly in the field of non-numerical evaluation scales, pre-analytical factors and matrix effects, point-of-care-testing, target values of control materials [116]. Recently, Haeckel has made an interesting proposal which could serve as an alternative to the second step in the hierarchy. It uses reference ranges, data from analytical methods and takes into account the rate of false-positive rates [47] and Krouwer has suggested to apply error grids [74].

1.4.2 Design of EQA schemes

An EQA scheme basically consists of preparing control samples, sending samples to the participants, and analyzing and reporting the participants' results. Initially, the EQA organizer selects an appropriate test material and determines which tests have to be determined by the participating laboratories. Control material should be tested for its homogeneity and stability before sending. Then, the EQA organizer sends the samples to the participating laboratories, often under controlled conditions, to preserve the stability of the parameters to be tested. Samples may be sent once a year and analyzed by the laboratories on a regular basis. Samples may also be sent at distinct periods and the laboratory analyzes the samples upon arrival in the laboratory. Samples are often shipped with an instruction form about their handling upon arrival, and a list giving an overview of parameters to be determined. Also, a description of the clinical case represented by the samples may be given. The laboratories then analyze the samples according to their standard routine procedures and return the results, preferably with interpretative comments, to the EQA organizer [43]. The results may be written down on a paper form or entered remotely in a web-based platform. After collection of the data, the EQA organizer applies statistical and graphical techniques to evaluate laboratories and the methods they applied. Usually, two reports are made. An individual report, focusing on the individual performance of the laboratory, is sent confidentially to each participating laboratory. Also, a report describing the performance of the laboratory in global terms and focusing on the performance of the analytical methods may be joined. Often, results of several rounds are compiled in an additional report, overviewing different clinical cases and describing a representative state of the art of testing by the participating laboratories. Also here a report with individual details for the laboratories may be provided confidentially. Laboratories with unacceptable performances may receive a warning letter inviting them to investigate their method and to undertake corrective measures.

1.4.3 Improvement of external quality assessment

Traditionally, External Quality Assessment has focused almost exclusively on monitoring and improving the analytical phase. However, it has been shown before that the analytical phase is the less error prone part of the Total Testing Process. EQA schemes should attempt to focus on the pre- and post analytical phases as well [107, 121]. Efforts have been made to build questionnaires about standard laboratory practices and questions about interpretation of examination results [82]. The Norwegian EQA organizer NOKLUS has developed Internet-based surveys to be sent to an extended group of clinical practitioners. Examples are given in literature for the interpretation of urine albumin [1], glucose test results [155] and warfarin monitoring [73]. In addition, the College of American Pathologists has organized a post-analytical survey about interpretation of elevated calcium results [56], and the IFCC Working Group Project “Laboratory Errors and Patient Safety” has issued a list of quality indicators that can be used to measure and evaluate laboratory testing in the different phases of the TTP [147]. Surveys investigating the interpretation of results tend to assess the post-post analytical phase instead of the post-analytical phase. Other enhancements are possible for each of the three phases of the Total Testing Process, as discussed in the following sections.

Assessing the pre-analytical process

Laboratories may be screened for behavior in case non-ideal samples or test request lists arrive. EQA organizers may serve themselves of simulated specimens, such as contaminated samples or samples with vague instructions [121, 177, 82]. It should be noted that in this phase, risk evaluation is different than the one commonly made. Errors appearing in this phase don’t have a continuous character and may make the end result completely wrong [185].

It is therefore important to evaluate in this phase the errors by frequency of occurrence.

Assessing the analytical process

EQA has historically been considered as an important tool in controlling analytical error, and also in this respect, improvements may be suggested. The use of commutable control material, i.e. material behaving in a similar way as native patient samples, is of interest to compare possible biasses between analytical methods and should be stimulated. In addition, the use of reference material is an added asset to inter-method comparison and the use of target values established by reference methods would be an ideal goal for all External Quality Assessment programmes. Also, when commutable control material is used, the comparison of result interpretations can be made without taking into account differences between methods. However, the mean of a subset of methods or techniques is also often used, mainly in cases of test specimens with so-called matrix effects, that do not affect patient specimens. In this case, true bias may be masked [131, 182] and laboratories exhibiting high precision should receive an optimal evaluation, disregarding the bias. The next chapter will deal with this issue in further detail. Furthermore, the use of an trueness-based rather than consensus-based evaluation creates room for improvement for, for example, glucose assays [197]. Another aspect which has not gained much attention yet is the monitoring of turnaround times. EQA organizers may use inquiries to obtain information about this issue [144].

Assessing the post-analytical process

The post-analytical process offers EQA organizers various opportunities of action, of which only a part has been explored so far. Laboratories may receive the description of a clinical case together with the samples they have to analyze. For the risk assessment of Down's syndrome for example, a risk can be calculated based on a combination of various measurements on the mother's blood and other clinical data, such as the mother's age, weight, and gestational age of the fetus. [177]. Another example is given by Sciacovelli [149], who has asked laboratories to formulate interpretative comments. Key phrases were identified in each comment according to the degree of abnormality found in the samples. Subsequently these phrases were allocated into key groups, and the consensus between laboratories in adding a brief interpretative comment to the patients' results was verified. Such a break down of interpretations into key phrases enhances a computerized processing of the data [78].

To mimic the Total Testing Process as much as possible, laboratories should be enabled to report to the EQA organizer in a way that is as similar as possible to their routine way of reporting. Nowadays many laboratories rely heavily on their Laboratory Information System (LIS). A LIS is a computerized platform enabling the capture of test results, storage and retrieval of adequate data, and a communication with other information systems. The international standard HL7 is a worldwide and generally accepted standard supporting communication between medical centers. It uses structured messages to transfer information between different systems.

Laboratory test results may be reported according to the Logical Observation Identifier Names and Codes (LOINC) [91] conventions. EQA organizers should include the possibility to communicate with their participants via HL7 and use LOINC codes to describe the tests they want to be performed. Using the power of LOINC, EQA organizers may also gain easily information about the applied reference ranges and interpretation of test results. Even more, laboratory reports can be made up based on HL7 messages and EQA organizers should strive to receive the data in computer-readable HL7-format and in human-readable laboratory test report format.

An automated way of reporting will not only enhance the communication between laboratories and EQA organizers, it will also avoid clerical errors while filling out web or paper forms, which are used nowadays. The latter are specific for EQA and are not representative for the Total Testing Process.

In this respect, it is interesting to read the comments on a question raised by Woods in 2004 [193] about the way EQA organizers handle obvious mistakes, such as a unit mistake or a wrongly placed decimal mark. Most responders said not to change any reported value [92, 190, 77, 187]. Only a minority mentioned unit mistakes as a reason why to change values [151, 129, 187]. In our opinion, values should never be changed without informing the laboratory and we suggest a 'correction phase' at the end of each EQA round. A laboratory may be shown an intermediate report, with a temporary evaluation of its results. z-scores (see sectio 1.5) and graphical representations (see Chapter 5), as shown on the final report, may be used here. Subsequently, laboratories may be invited to screen their results.

Data which are the result of an error are supposed to have high z-scores, or to be shown as an outlying observation on the graphical representations, and should attract the attention of the reader. It would then be advisable to give the opportunity to the laboratories to comment on their own results. Exceeding values may be flagged as an error, and specifications, such as pre-analytical, analytical or post-analytical error may be specified. If possible, the laboratory should have the possibility to respond the true analytical result of the test.

This approach will not only enable EQA organizers to use cleaner data in calculating their measures of performance, it will also be more informative to the laboratories. Moreover, summarizing statistics about the source of the error reflect the laboratories' performance in the light of pre-, middle- and post-analytical error, and may be of interest to the participating laboratories. Lastly, only corrected values should be used for any kind of calculations. EQA organizers should not be blinded by the robustness of their methods. Robust methods for estimating mean and variance are less sensitive to outlying results than standard methods, they are still sensitive to it.

1.5 Statistical analysis of EQA data

1.5.1 The z-score concept

Ideally, every laboratory participating in an External Quality Assessment programme should report a value which is close to the value which was expected to be reported. Every reported value which is 'too far away' from the expected value looks peculiar, and it is the principle of the evaluation procedure to find those values which are too far from the expected value. This logic forms the base of z-scores, which have become a way to assess the quality of clinical laboratories by classifying them on a common continuous scale and flagging those with unacceptable results. Besides, these scores can be interpreted similarly to those of the internal quality procedures. They are based on a measure of center and scale of the distribution of the results, in which the difference from the center is expressed as a multiple of the measure of scale:

$$\mathrm{z-score} = \frac{\mathrm{x} - \mu}{\mathrm{s}}$$

where x is the individual value for a laboratory, μ is the assigned value and s is the standard deviation used to evaluate the laboratories.

There is a common agreement to flag z-scores with absolute values larger than 3, requesting an action signal from the laboratory. Z-scores with absolute values larger than 2 and smaller than 3 are regarded as a warning signal, and those with absolute values smaller than 2 as within acceptable limits [2]. For a Normal distribution without any outliers, there is a chance of about 1 on 1000 to have an unacceptable z-score. Z-scores obtained for several samples may be combined and the frequency of exceeding a limit, 3 or lower, may be used as a long-term evaluation tool. Therefore, z-scores convey a different kind of information and are more flexible than outlier detection techniques, which search only for discordant values.

The estimation of location and scale is of primary importance for obtaining z-scores. In some cases, the estimates are known or fixed beforehand as derived by a reference method and/or a fit for purpose standard deviation [3]. When mean (center) and standard deviation (scale) are calculated from the sample values, however, they can be heavily influenced by outliers. For example, when the sample size n<12, outlying results influence the estimates in such a way, that z-scores are never larger than 3 [4]. Estimation of the center and scale by avoiding the influence of outliers can be done by two distinct types of approaches [3,5].

The outlier-based approaches first exclude outlying values, calculate the classical mean and standard deviation on the remaining data, and then compute z-scores for all values, including those previously excluded. Other approaches use robust statistics, which obtain measures of location and scale which are less influenced by outliers. Healy introduced robust statistics in the field of laboratory medicine [6] while Rocke described its use in EQA for the first time in 1983 [7]. Today, robust statistics are popular in the domain of EQA, as confirmed by the ISO recommendations for calculating z-scores [2]. In EQA, standard deviations are not solely used for detecting laboratories that reported "out-of-range" outlying values, but also serve as quality indicators for analytical methods, in particular with respect to trueness and precision. Here, the uncertainty and bias of variability estimators themselves are of importance.

Finally, it is worth mentioning that the distribution of EQA data is often assumed to be symmetric and possibly contaminated with few outlying results and most statistical tests in EQA programmes assume a contaminated Normal distribution. Some authors [3,4], however, claim that the distribution, even in the absence of outliers, may be leptokurtic, i.e. exhibiting heavier tails than the Normal distribution. Thus, comparing the two types of estimating approaches for the specific classes of symmetric and unimodal distributions, represented by a Normal and a Student's t-distribution may be of importance The purpose of this study [26] was to answer three simple questions:

(1) what is the false positive rate of z-scores estimation methods in non-contaminated samples from the Normal and Student's t distributions ?

(2) what is the true positive rate of z-score estimation methods in contaminated samples from the same distribution ?

(3) what is the accuracy and precision of the different variability estimators of the Normal distribution ?

1.5.2 Estimating z-scores

Approach 1: Grubbs test

Although mainly used for sample sizes larger than 20, the Grubbs test [44] may theoretically be used for smaller samples. It starts with calculating the quantity G:

$$G = \frac{\max\limits_{1 \leq i \leq n} |x_i - \overline{x}|}{s}$$

Now, if

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{Q_t^2(1-\frac{\alpha}{2n}; n-2)}{n-2+Q_t^2(1-\frac{\alpha}{2n}; n-2)}}$$

where $Q_t(1-\frac{\alpha}{2n}; n-2)$ is the upper $\alpha/2n$-quantile of the Student t-distribution on n-2 degrees of freedom, the most extreme value will be left out of the data set and the Grubbs test will be applied again. This process is repeated until the above inequality does no longer hold. Then, the average and standard deviation are calculated on the reduced dataset and are used to calculate z-scores for all the data, uncluding those excluded in the iterative process. Another possibility exists of using a blocking procedure [139, 64]. However, this iterative procedure was preferred since the number of data in this experiment, and hence, the number of possible outliers, are small.

Approach 2: Dixon test

This approach is based on an article published in 1950 by Dixon [30], who developed several approaches to find outliers for sample sizes as small as n=3.

The method is based on the calculation of ranges between lowest and highest values of the samples, and subranges between the most extreme values from the samples. It starts from an ordered sample $x_1 \leq x_2 \leq \ldots \leq x_{n-1} \leq x_n$ and calculates the ranges depending on the sample size:

For sample sizes ranging from 3 to 7, the following ratio is calculated:

$$r = \frac{x_2 - x_1}{x_n - x_1} \text{ for outliers at the lower side}$$

$$r = \frac{x_n - x_{n-1}}{x_n - x_1} \text{ for outliers at the upper side of the sample}$$

For sample sizes ranging from 8 to 10, the ratio is slightly modified:

$$r = \frac{x_2 - x_1}{x_{n-1} - x_1} \text{ for outliers at the lower side}$$

$$r = \frac{x_n - x_{n-1}}{x_n - x_2} \text{ for outliers at the upper side of the sample}$$

For sample sizes ranging from 11 to 13, we have:

$$r = \frac{x_3 - x_1}{x_{n-1} - x_1} \text{ for outliers at the lower side}$$

$$r = \frac{x_n - x_{n-2}}{x_n - x_2} \text{ for outliers at the upper side of the sample}$$

For samples of larger size, the calculation of r becomes:

$$r = \frac{x_3 - x_1}{x_{n-2} - x_1} \text{ for outliers at the lower side}$$

$$r = \frac{x_n - x_{n-2}}{x_n - x_3} \text{ for outliers at the upper side of the sample}$$

Dixon has described in 1951 [31] the distribution of r, which can be used to build a hypothesis test around it. A low P-value of the hypothesis test then indicates that the sample contains an outlier. In this case, the most extreme value is removed and the test is applied to the reminder of the data till there is no significant proof of an outlier any more. The classical average and standard deviation are calculated on the reminder of the data and are applied to all the data, also the values earlier detected as outliers.

Approach 3: Tukey's robust method
This approach is based on the assumption that the 25th percentile (P_{25}) and 75th percentile (P_{75}) are generally not influenced by outliers, so that the standard deviation can be calculated by the following formula:

$$\mathrm{s} = 0.7431 \times \mathrm{IQR}$$

where $\mathrm{IQR} = \mathrm{P}_{75} - \mathrm{P}_{25}$ is the so-called interquartile range and the factor 0.7431 is the reciprocal of the difference between the 25th and 75th percentile of a standard Normal distribution.

Approach 4: Qn estimator
The Qn estimator [140] is given by sorting all pairwise absolute differences of sample values, namely $|\mathrm{x_i} - \mathrm{y_j}|$ (i<j; i=1,n-1; j=2,n) and taking the kth order statistic from the ordered series. If we denote by $\left\{|\mathrm{x_i} - \mathrm{y_i}|\,;\mathrm{i} < \mathrm{j}\right\}_{(\mathrm{k})}$ this order statistic, then

$$\mathrm{Qn} = \mathrm{D}\left\{|\mathrm{x_i} - \mathrm{y_i}|\,;\mathrm{i} < \mathrm{j}\right\}_{(\mathrm{k})}$$

where D is a constant factor and equal to $1/\left[\sqrt{2}\Phi^{-1}\left(\frac{5}{8}\right)\right]$, with Φ^{-1} the inverse cumulative distribution function of the standard Normal distribution, and $\mathrm{k}=\mathrm{C}_2^{\mathrm{h}} \approx \mathrm{C}_2^{\mathrm{n}}/4$, where $\mathrm{h}=\frac{\mathrm{n}}{2}+1$, is close to half the sample size.

Approach 5: ISO 13528
The ISO 13528 standard's algorithm for calculating mean an standard deviation is as a robust algorithm to be used in External Quality Assessment (see Algorithm A in Appendix C of the standard). It makes use of the principle of Winsorization and starts with the calculation of the median of the data:

$$\mathrm{x}^* = \mathrm{median}$$

followed by a first robust estimate of the standard deviation:

$$\mathrm{s}^* = 1.483\,\mathrm{median\ of}\ \{|\mathrm{x_i} - \mathrm{x}^*|\,;\ \mathrm{i} = 1, \mathrm{n}\}$$

Next, the quantity δ is calculated

$$\delta = 1.5\mathrm{s}^*$$

and for every $\mathrm{x_i}$, an alternative value $\mathrm{x_i^*}$ is attributed as follows (i=1,n):

$$\mathrm{x_i^*} = \begin{cases} \mathrm{x}^* - \delta & \text{if } \mathrm{x_i} < \mathrm{x}^* - \delta \\ \mathrm{x}^* + \delta & \text{if } \mathrm{x_i} > \mathrm{x}^* - \delta \\ \mathrm{x_i} & \text{otherwise} \end{cases}$$

Then, new values for x^* and s^* are calculated by the equation

$$\mathrm{x}^* = \frac{\sum \mathrm{x_i^*}}{\mathrm{n}}$$

and

$$\mathrm{s}^* = 1.134\sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i^*} - \mathrm{x}^*)^2}{\mathrm{n} - 1}}$$

The new robust estimates of x^* and s^* are used to calculate new values of $\mathrm{x_i^*}$ and the iteration is continued till convergence.

1.5.3 Simulation study

A total of 1000 random samples was generated from a Normal distribution with a population mean and standard deviation arbitrarily set at μ=10 and σ=0.5. The data were generated for sample sizes from n=3 to 20. Subsequently, to obtain data from a leptokurtic distribution, a similar simulation was performed using a Student's t-distribution with 5 degrees of freedom. Only samples for which all values were within the interval [μ-3σ,μ+3σ] were withheld. Next, the samples were contaminated by adding an outlier at $\mu+3\sigma$, at $\mu+5\sigma$ and at $\mu+7\sigma$ separately, resulting in a sample of size n+1. Samples of size of n=3 without added outliers were not taken into account. z-scores were calculated on each sample following the five different approaches. The Grubbs and Dixon test were applied with an α of 0.05.

The ability of various approaches of flagging outliers when they exist and of not flagging them when they don't exist can as well be assessed in a way similar to the evaluation of diagnostic tests: the first question addresses the true positive rate and the second question addresses the true negative rate.

For this purpose, Negative Predictive Value (NPV) and Positive Predictive Value (PPV) were calculated for each approach by modifying each algorithm in letting a parameter vary that may be changed to increase or decrease the number of z-scores above 3. For the Grubbs and Dixon tests-based approaches, the P-value for which outliers are excluded (α) was changed. For the Qn approach, D was to be changed and for the Tukey approach, the factor that is multiplied with the IQR. Lower values of these two factors result in lower standard deviations, higher z-scores and a higher z-citation rate. For the ISO-13528 approach, δ was chosen to be changed. NPV and PPV for each of the different values of the varying parameter were recorded and graphically displayed.

At last, for each simulated data series of samples generated from a Normal distribution, the variability estimator obtained by every approach was recorded and its mean and standard error calculated.

1.5.4 Results

False outlier rates - The false outlier rates are visualized in Figure 1.2. Among all approaches, the Tukey method showed the most distinctive behaviour; while all false outlier rates were below 15% for the Normal distribution and below 30 % for the Student's t-distribution, Tukey's approach

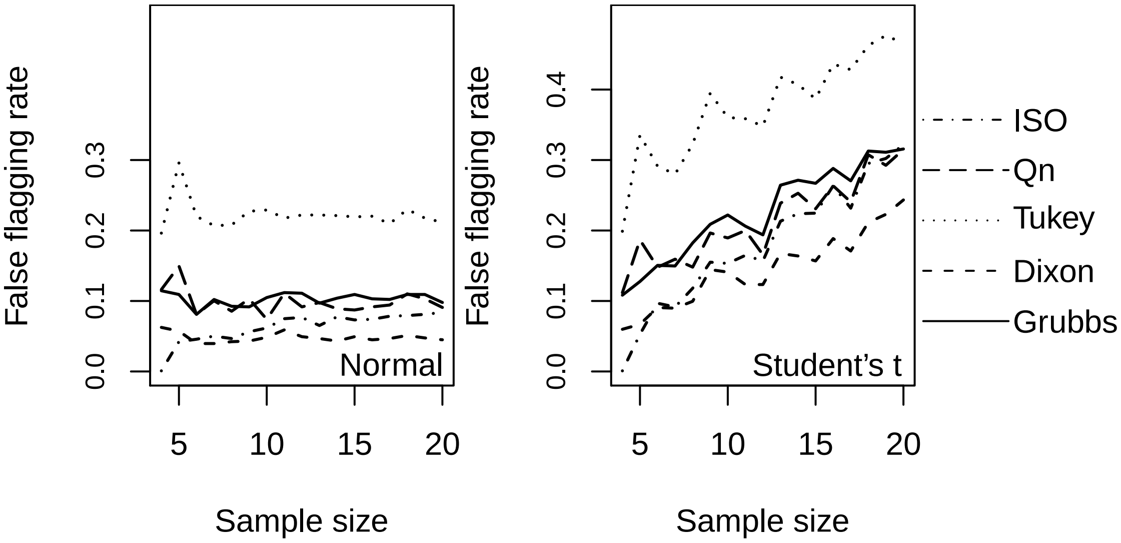

Figure 1.2 False flagging rates for the five different approaches, for different sample sizes. Left graph shows results for the Normal distribution, right graph those for the Student's t distribution with 5 degrees of freedom. Results are based on a simulation with 1000 samples.

had for almost all sample sizes a rate above 20%. The Dixon and ISO approaches showed the lowest false outlier rates. In addition, it is seen that for samples of size 6 or larger, false outlier rates didn't change much for the Normal distribution, except for the ISO approach. However, all increased with increasing sample size for the Student's t-distribution.

True outlier rate - The flagging rates when adding an outlier at varies distances are shown in Figure 1.3. For all outlier distances, differences between the different approaches were similar for the Normal and Student's t-distribution. For outliers at µ+3σ, none of the approaches was able to flag the outliers in more than half of the cases for all sample sizes. The Tukey approach had the highest performance, reaching a flagging rate of nearly 50% as soon as the sample size was 6 or larger. The other approaches had much weaker performance; the ISO approach had flagging rates below 10% for very small samples. All other approaches exhibited outlier finding rates of roughly 10-30%.

For outliers at µ+5σ, a strong improvement of the Grubbs test-based algorithm is obvious, outperforming the Tukey approach and reaching an outlier finding rate near 100% for sample sizes of 10 or larger. Also here, the ISO approach has a very weak performance for very small sample sizes. The outlier finding rate increases strongly with increasing sample size for all approaches.

The results point to a clear improvement of flagging rates for all approaches with increasing sample size and outlier disance, with a probability of detection close to 100% for outliers at µ+7σ. The ISO and Dixon approaches, however, still had a weak performance for very small sample sizes.

Negative and Positive Predictive Value - The results of the NPV and PPV for sample size n=6 is shown in Figure 1.4 and for sample size n=8 in Figure 1.5. The perfect approach, which would flag no z-scores larger than 3 in case they don't exist (negative prediction) and would flag them all in case they would exist (positive prediction), would show a curve made up by a vertical line equal to the Y-axis and a horizontal line which intersects the Y-axis at the value 1. The further the curve of an approach departs from the perfect curve, the worse its performance.

For outliers at µ+3σ, curves were located far from the ideal line and NPV and PPV for none of the approaches reached high levels: only a combination of positive and negative predicted values of about 60% was feasible, and although the Grubbs approach tended to perform better, there was not much difference between the approaches. For increasing outlier distance, however, it is seen that the curves tend to close the curve of the perfect approach. The Qn approach consistently performed the worst. The outlier searching algorithms showed a slightly better performance, mainly for outliers at moderate distance from the center (µ+5σ). There was almost no difference between the results of the data generated from the Normal or from the Student's t-distribution. A similar trend was seen for a sample size n=8. All algorithms exhibited a weak performance for outliers at µ+3σ. For outliers at

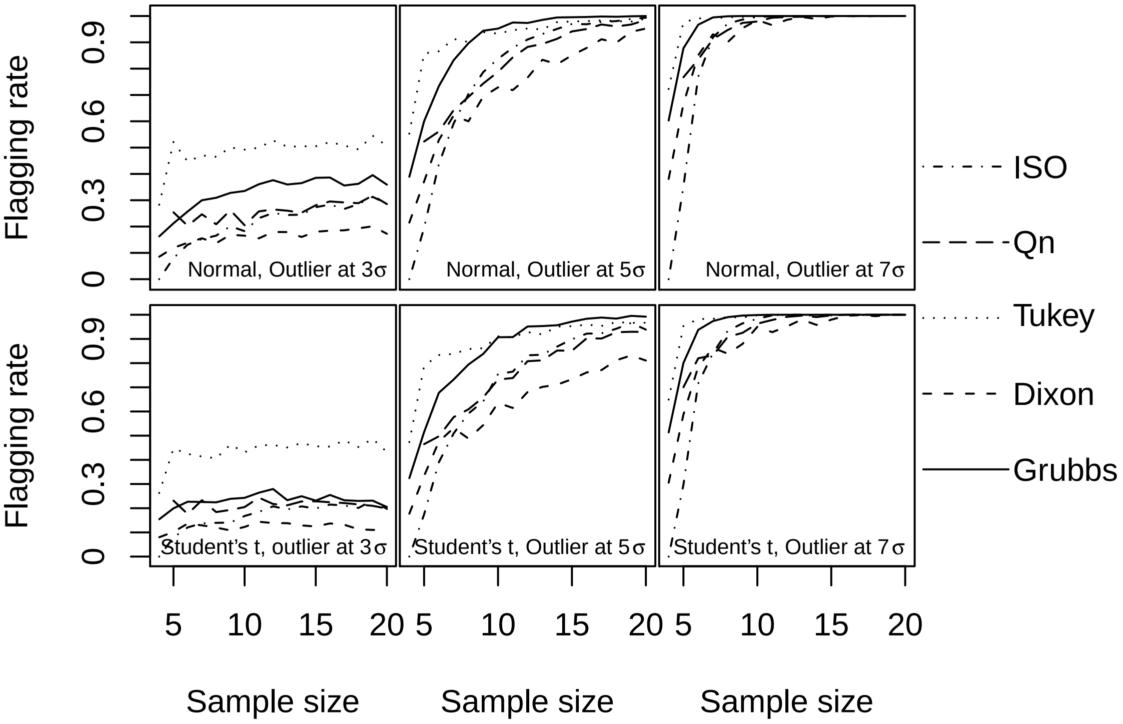

Figure 1.3 Flagging rates for the five different approaches with an outlier at various distances from the center, for different sample sizes. Upper graphs show results from the Normal distribution, lower graphs those from the Student's t distribution with 5 degrees of freedom. Results are based on a simulation with 1000 samples.

μ+5σ, however, the Grubbs approach performed better than the other approaches. This difference became less clear for more distant outliers, where all approaches showed almost perfect positive and negative predictive values.

Focussing on the Grubbs approach, the search for the optimal P-value for excluding outliers (α) was made for different combinations of sample size and outlier distance. The optimal α decreased when outliers became more distant and with increasing sample size. For outliers at small distance from the distribution, the most optimal α was 0.2 for all sample sizes. This value decreased when the outliers was further away from the distribution (0.02-0.1 for outlier at μ+5σ, 0.007-0.06 for outlier at μ+7σ) (See Table 1.1).

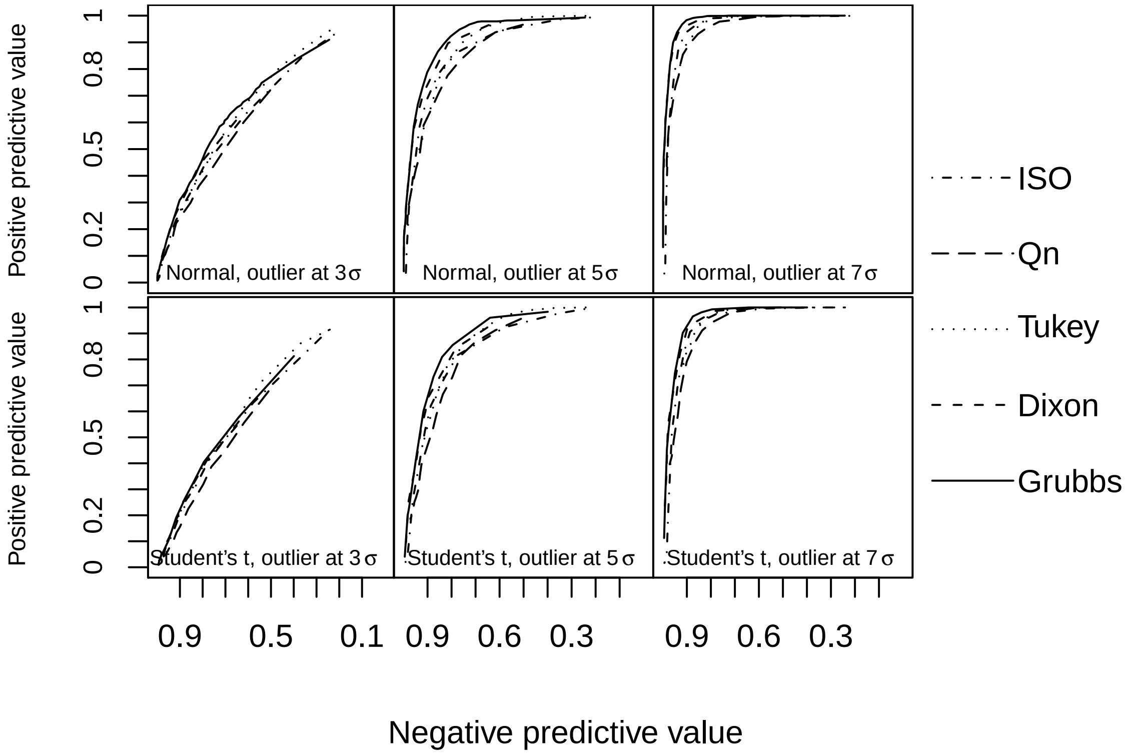

Figure 1.4 NPV and PPV for the five different approaches at different outlier distances, sample size n=6. Upper graphs show results from the Normal distribution, lower graphs those from the Student's t distribution with 5 degrees of freedom. Results are based on a simulation with 1000 samples.

Variability and bias of standard deviation - The results are depicted in Figure 1.6. In absence of outliers, Tukey and outlier search-based approaches showed a higher distance between the estimated and actual population mean of 0.5, consistently underestimating the standard error. The trueness and precision of every estimator becomes worse when an outlier is present at µ+3σ. Precision is more or less equal for all approaches. Trueness however exhibits major differences between the approaches. The Tukey and Grubbs approaches suffer the least from lack of trueness. Also here, trueness and precision improve with increasing sample size.

Precision and trueness for outliers at µ+5σ show that the Qn and ISO estimator consistently overestimate the standard deviation. Only the Grubbs

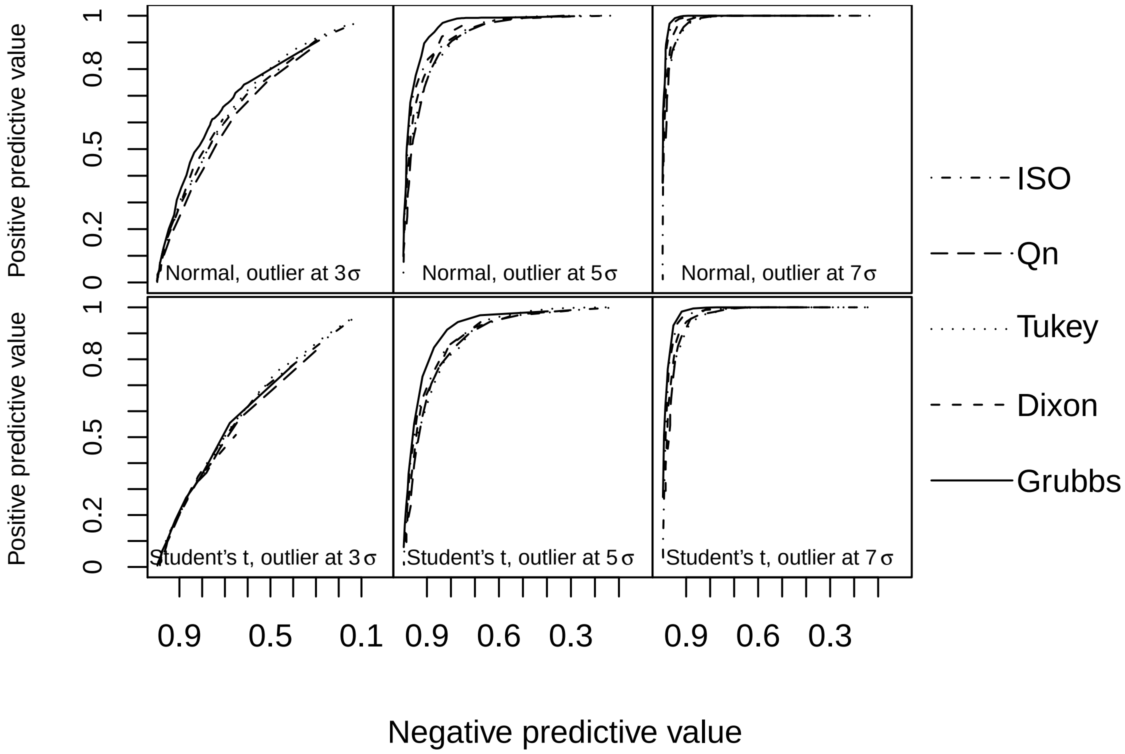

Figure 1.5 NPV and PPV for the five different approaches at different outlier distances, sample size n=8. Upper graphs show results from the Normal distribution, lower graphs those from the Student's t distribution with 5 degrees of freedom. Results are based on a simulation with 1000 samples.

Table 1.1 Optimal values of α at which outliers should be searched for, for having best combination of positive predictive value and negative predictive value for the Grubbs approach to calculate z-scores.

n		Outlier	
	μ+3σ	μ+5σ	μ+7σ
5	0.20	0.10	0.06
6	0.20	0.08	0.04
7	0.20	0.08	0.04
8	0.20	0.06	0.02
9	0.20	0.04	0.02
10	0.20	0.04	0.02
11	0.20	0.04	0.009
12	0.20	0.04	0.007
20	0.20	0.02	0.007

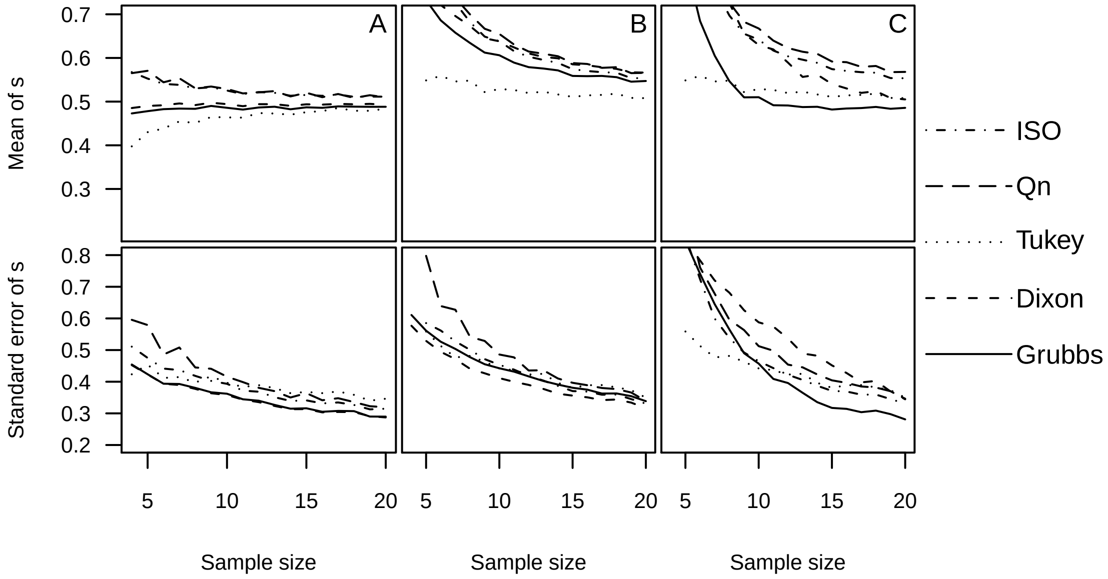

Figure 1.6 Trueness (upper graphs; expressed as s and compared with true value of 0.5) and precision (lower graphs, expressed as standard error of s) of estimates of standard deviation obtained by the different approaches for the Normal distribution. A: no outliers added; B: outlier at μ+3σ; B: outlier at μ+5σ. Results are based on a simulation with 1000 samples.

and Tukey approaches are able to give estimates of the standard deviation with high truness, the former starting from an sample size of 8, the latter for all sample sizes. Also, the Grubbs approach outperforms the other approaches in terms of precision.

The difference between the approaches becomes even clearer when outliers at μ+7σ are considered. The precision of the Grubbs approach outperforms all the other approaches, also at small sample size. Regarding at the trueness, we can see that the outlier-search approaches, together with the Tukey approach, have a higher trueness than the other approaches.

1.6 Conclusion

The world of the clinical laboratory has undergone major changes in the last decades. Today it is considered as a knowledge center, not only for producing but also explaining analytical results to clinicians. It has become a prerequisite for diagnostic and therapeutic decisions. Because of its crucial role, and definitely after the release of the Institue of Medicine's report 'To Err is Human: building a Safer Health System', laboratories must attein high standards to avoid errors. For avoiding and discerning the possible causes of errors, the implementation of a quality management system has become of utmost importance.

External Quality Assessment is seen as a cornerstone of quality management. It is the discipline in which results obtained for the same sample by different laboratories are compared with each other. Historically, the analytical phase has received most attention in EQA, in the sense that many authors still try to explain deviating results in an EQA round by high bias or variability, or a combination of both. EQA organizers, however, should not solely use flaws in the analytical process as possible explanations of spurious results. Errors may have taken place in the pre and/or post analytical phase as well. In addition, EQA organizers should strive towards assessing pre- and post-analytical phases, since investigations about errors have shown that the pre- and post analytical phase are the most error-prone. Questionnaires are a useful tool to assess both phases, and specific surveys can be set up, for example by using specific sample material for assessing the pre-analytical phase. EQA organizers could however assess the post-analytical aspect of laboratory medicine by receiving the laboratory results and their interpretation in the same way as clinicians receive them. In fact, they should strive towards closing the gap between the usual data flow in the routine laboratory and the typical data flow for an EQA survey. Technical solutions exist, preferably under the form of HL7. In addition, EQA organizers have to be aware of a small fraction of data that are heavily deviating. Enabling the laboratories to rectify these data in a documented way, is one of the keys to better statistical estimates of bias and variability.

There are various approaches for calculating mean and variabilty, and although all of them perform well for larger sample sizes and outliers at a distance of several standard deviations from the mean, there are some differences at smaller sample sizes. In general, the Tukey approach will indicate more often z-scores above 3 than other approaches, whether outliers are present are not, and hence overestimate the frequency of outliers. In comparison with other approaches however, it yields, in the presence of an outlier at small distance, the least biased estimate of variability. If this approach is used, we want to recommend a minimal sample size of 6 since smaller sample sizes have a combination of too high false outlier rates or too low true outlier rates. The Qn and ISO approaches perform in a similar way, although the latter has a very low true outlier finding rate at very small sample sizes. They behave as good as unbiased estimates of the variability when no outliers are present, but overestimate it heavily as soon as an outlier at small distance is present. Even for sample sizes of 20, they overestimate the variability in presence of outliers. We would not recommend to use these approaches for sample sizes smaller than 10. The outlier-search approaches, in particular the Grubbs approach, tend to perform better when considering true and false outlier finding rates together. In addition, at least for sample sizes of 6 and larger, the estimator of variability obtained by the Grubbs approach is more stable and less biased than the other approaches in the presence of outliers at far distance. For these reasons, we would recommend the Grubbs approach for calculating z-scores. As a rule of thumb, we would propose to use an α of 0.05 for all sample sizes. If different values of α can be used, we would propose to let α be 0.1 for samples of size 10 and smaller, and 0.05 for samples of larger size.

CHAPTER 2

Matrix effects, EQA schemes and standardization between methods

2.1 Introduction

One of the major problems in laboratory medicine today is the poor comparability of results of patient samples that were obtained in different laboratories using different analytical methods. Differences between results obtained for the same patient may jeopardize correct clinical decisions [114, 22]. Patients are frequently treated by a team of physicians rather than one, often extending across several health care disciplines and making use of information from various laboratories. Comparability of laboratory results is also important for tests with regionally or nationally established decision limits, or for tests used for monitoring patients over a long time period [164, 99, 114]. In addition, accurate results also allow laboratory data to be collected and mined from different sources in order to identify public health needs or monitor public health programmes [180, 181]. As a consequence, calibration and harmonization of results from different analyzers, both between and within laboratories, and the continuity of such harmonization in time are of great importance [8, 65].

The key to comparable results is standardization. Its goal is to assure that results from measurements in patient samples are accurate, independent of the measurement procedure used or the location and time of testing. In statistical terms, we speak about eradicating bias between methods. To achieve standardization, the approach of traceability may be adopted from the field of analytical chemistry. It is based on the reference measurement system, in which, starting from reference material or standards that are nationally or internationally recognized, a reliable transfer is provided towards routine methods using metrological traceability [113, 39, 33]. The importance of the metrological principles has been recognized by the International Organization for Standardization (ISO), which has written down rules to ensure standardization in ISO 17511 [58] and 18153 [114, 59]. Later, the European Commission has issued the Directive on In Vitro Diagnostic Medical Devices (IVD) 98/79/EC [36]. It requires the clinical diagnostic industry to document the metrologic traceability of in vitro diagnostic systems towards standards supported by the international system SI [176, 8]. However, manufacturers have prepared their own calibrators, which are not always available for other manufacturers. Often, standards of higher order are stabilized, for example by lyophilization, and lack similarity with human samples [136]. This can lead to a disagreement between results from different commercial assays [113, 115, 114]. Further, since the IVD directive doesn't recognize internationally recognized reference materials, manufacturers have to select their own reference systems, which may also cause disagreement between results [6]. For this reason, the IVD directive also foresees a role for EQA schemes in assessing agreement between routine methods, also known as post-market vigilance [175]. In addition, a recenlty published roadmap for harmonization between laboratories also underlines the role of EQA using commutable samples [103]. Although EQA programmes are by nature excellently placed for this job, not every EQA survey can serve for post-market vigilance due to the possible presence of matrix effects in the utilized sample material.

2.2 Matrix effects

EQA organizers often use sample material specifically prepared to ease transportation and storage, and that can be produced at a relatively low cost, exhibits a low vial to vial variability and can be used to determine a wide range of parameters on the same specimen [98, 52]. For this, samples are prepared using a pool of human sera or plasma, are often lyophilized, and stabilizers and other substances may have been added. Several analytes may also have been added by spiking to obtain a wide range of values that can be determined on one single sample. These preparation steps, however, may have adverse effects on the physicochemical properties of the samples. For example, lyophilization irreversibly denaturates lipoproteins, causes modifications in viscosity, and increases turbidity and alters the pH and surface tension [39]. When the analyte is spiked using material of non-human origin, complexes may be formed jeopardizing a correct measurement [40, 52]. Protein complexes may be modified during isolation from human sources and can produce a different measurement signal than expected for native forms of the analyte [99]. Analytical techniques based on immunological reactions, such as immunoassays, are sensitive to factors influencing antigen-antibody reactions, such as a specificity of antibodies after preparation [33, 41, 196]. As a consequence, samples may lack similarity with genuine samples analyzed in the laboratory [175] and several methods may obtain different results for a certain parameter, a characteristic commonly called “matrix effects” or “non-commutability”.

The term “commutability” was first used to describe the ability of a reference or control material to exhibit properties comparable to the properties of authentic clinical samples when analyzed by different analytical methods. This description is now more generally defined as the equivalence of the analytical results of different measurement procedures for a reference material and for representative samples from healthy and diseased individuals. The ISO defines commutability as the degree to which a material yields the same numerical relationships between results of measurements by a given set of measurement procedures. Some authors define it in a more method-specific narrow way: it means that a reference or control material will behave in the same way as a genuine human sample. This implies that a sample may be commutable for one specific method, but not for another [180, 99, 179, 98, 136, 130]. Matrix effects are defined as “a bias or difference caused by a sample property other than the level of the substance or property that is intended to be measured”. It includes physicochemical, mechanistic and analytical interferences and substance isoforms [75].

In general, we can say that matrix effects in a processed material are caused by an altered matrix which would not be expected to occur in typical authentic clinical specimens and thus represents a difference between the EQA sample material and authentic clinical specimens. In most cases, the presence and magnitude of a matrix bias is unknown [98]. Matrix effects are somewhat unpredictable and their frequency may vary largely in the different experiments. For these reasons, Miller stated that prepared samples are not suitable for field-based postmarketing assessments of standardization [102, 100]. Moreover, recently, he stated that quality control samples should not be used to verify the consistency of results for patient samples when a new reagent lot is used [101].

2.3 Solutions for avoiding matrix effects

Two different approaches may be adopted to cope with matrix effects. First, EQA organizers may attempt to produce commutable samples. Secondly, when commutability cannot be assured, laboratories should be partitioned into so-called peer groups. A peer group is a group of laboratories using the same or similar analytical methodology for the determination of a certain parameter, grouped in such a way that the group is free of matrix-dependent bias.

2.3.1 Improving sample quality

One approach to solving the non-commutability problem is to use authentic clinical specimens or specimen pools for EQA. This approach has been effective when adequate specimens can be obtained but has been limited when multiple analytes, large quantities or long transportation distances are required. A discussion paper from EQA working group B on target values in EQA [172] recommended two different approaches for EQA. Fresh-frozen native patient samples were recommended for method assessment using reference method target values. Non-commutable processed sera could be used for participant assessment, in which case reference method values were informational only.

Careful preparation of a fresh-frozen off-the-clot serum pool, or other native clinical specimen, is critical to the validity of the assumption that the material is commutable. A consensus guideline for preparation of a fresh-frozen off-the-clot serum pool is CLSI document C37-A [24] which specifically addresses cholesterol but is applicable to most serum measurands [98]. Stöckl and Thienpont [176, 160] have proved that single individual donations can be used as EQA material; the only differences these samples show with samples usually analyzed in routine labs are filtering and, possibly, freezing. The main limitation for this kind of surveys is the volume of sample material to obtain. However, knowing that modern clinical analyzers need only small volumes of sample, using aliquots of 0.5 mL, or even only 0.3 mL, about 600 samples could be used from one single donation [160].

A number of EQA programmes are run in several countries that use commutable samples prepared from freshly collected and minimally processed human samples. One of the oldest is the Glycohemoglobin Survey from the College of American Pathologists in the USA. Hybrid approaches, in which processed and authentic clinical specimens are used in the same EQA round, represent a practical step forward to add value for individual participants and for method manufacturers. Commutability provides a point of comparison for laboratories and manufactures with respect to the IVD directive for a certain set of parameters, and the prepared sample may be used for a screening on a larger magnitude of parameters. [98].

2.3.2 Partitioning of results into peer groups

When matrix effects cannot be avoided, the most common procedure is to group the participants according to their method into so-called “peer groups”. The latter are formed per parameter and gather the laboratories that use the same or similar methodology. It is for example common to group the laboratories according to the manufacturer of the material used for analyzing a certain parameter. They are likely to achieve the same analytical result and their reported values should follow a unimodal distribution, eventually contaminated by a low fraction of outliers. Z-scores may be calculated using the algorithms described in Chapter 1. It should be clear here that the target value, taken by the median or the mean after outliers are excluded, may be different for each distinct peer group and hence the conclusions are less far-reaching than those obtained by fresh samples. Peer group evaluation does confirm that a laboratory is applying a technology correctly and does measure the uniformity of the manufacturer’s field method calibration process among a group of users [100, 98]. It does however not bring any information about a possible bias between methods for routine samples.

Homogeneity is a major prerequisite of peer groups. Lack of homogeneity of the peer group inflates the standard deviation and decreases the flagging rate by masking results which should be flagged. The creation of peer groups is a compromise between reducing uncertainty about variability estimates by taking peer groups as large as possible, and reducing inflated standard deviations by taking peer groups as homogeneous as possible. These two objectives become antagonistic when values obtained with slightly different methodology is used and the question arises whether these values should be partitioned in one or two peer groups. The increase of the standard deviation when two heterogeneous groups are merged may be the key to the answer.

If one heterogeneous group consists of two different homogeneous groups, the mean and standard deviation will change and not reflect any more the mean and standard deviation from the original groups. To calculate the effect of heterogeneity, we can start with a sample data of size n and mean $\bar{x}_1$, in which a proportion p has shifted by a value δ to form a new subgroup with mean $\bar{x}_2 = \bar{x}_1 + \delta$. The mean of the heterogeneous distribution becomes $\bar{x} = \bar{x}_1 + p\delta$. Assuming that the original standard deviation of the data series, s_1, remains the same for the two subgroups, the calculation of the standard deviation of the heterogenous group then becomes

$$
\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \sum_{i=1}^{(1-p)n} (x_i - \bar{x}_1 - p\delta)^2 + \sum_{i=(1-p)n+1}^{n} (x_i - \bar{x}_2 - (p-1)\delta)^2 \\
&= \sum_{i=1}^{(1-p)n} (x - \bar{x}_1)^2 - \sum_{i=1}^{(1-p)n} 2p\delta(x_i - \bar{x}_1) + \sum_{i=1}^{(1-p)n} (p\delta)^2 + \\
&\quad \sum_{i=(1-p)n+1}^{n} (x_i - \bar{x}_2)^2 - \sum_{i=(1-p)n+1}^{n} 2p\delta(x_i - \bar{x}_2) + \sum_{i=(1-p)n+1}^{n} [(p-1)\delta]^2 \\
&= \Big[(1-p)n - 1)s_1^2 + (1-p)n(p\delta)^2 + (pn-1)s_1^2 + pn((p-1)\delta\Big]^2 \\
&= (n-2)s_1^2 - p^2 n\delta^2 + pn\delta^2
\end{aligned}
$$

Thus

$$
s = \sqrt{\frac{(n-2)s_1^2 + pn\delta^2 - p^2 n\delta^2}{n-1}}
$$

The ratio $\frac{n-2}{n-1}$ appears in the formula because the standard deviation is based on the estimation of two different averages. For n large this ratio approaches 1 and then we have approximately

$$s = \sqrt{s_1^2 + p\delta^2(1-p)}$$

In case half of the data show an upwards shift of δ, p=0.5 and the formula becomes:

$$s = \sqrt{s_1^2 + 0.25\delta^2}$$

This formula is visualized in Figure 2.1.

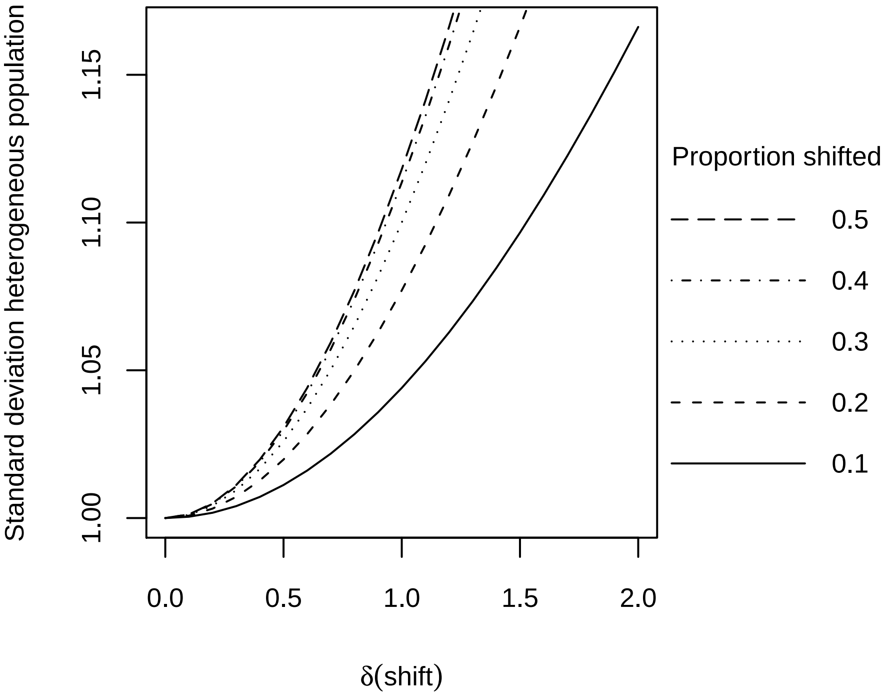

Figure 2.1 Behavior of standard deviation when a certain proportion of the data exhibits a shift (δ). Lines are calculated based on an sample of n=20.

It is clear that the standard deviation increases more than proportionally with increasing shift. Also remark that the standard deviation increases faster when the proportion of shifted data is larger.

This model can be employed to investigate the influence of combining two possibly heterogeneous groups into one peer group. Assume for example that the original standard deviation of both groups is 1 and a shift in the standard deviation of 5 % is allowed. The maximal standard deviation inflation would be reached by a shift of 50 % of the data, and the maximum shift can then be calculated as follows:

$$1.05 > \sqrt{1 + 0.25\delta^2}$$

or δ<0.64. In other words, as long as the bias between the two group means is smaller than 0.64 and their standard deviation is 1, the shift in standard deviation will be below the acceptable limit of 5 %.

Detecting heterogeneity in a peer group

It is interesting to know when, based on a number of samples used in EQA rounds, the difference between two group means is smaller than the maximal allowable shift. A Multivariate Analysis of Variance (MANOVA) test may be of interest. In classical analysis of variance, one checks whether one or more groups differ with respect to a given variable. MANOVA does the same, but checks whether one or more groups differ from each other for a multivariate set of variables. In the case of EQA, responses obtained for a series of samples for one particular parameter may serve as a multivariate set, and two different methods may serve as the explanatory variable.

A MANOVA F-test, like many other hypothesis tests, yields clear results when the null hypothesis is rejected: there is enough evidence to state that there is a difference between the groups, and the chance of being mistaken is very small. Remark that, when enough data is available, the P-value of the test will be significant, even if the difference between the groups is small, also when it is smaller than the maximum allowable difference. When, on the other hand, the P-value is larger than a predefined threshold value α (often taken 0.05), one has to decide that there is no evidence of a difference. Saying that there is no difference would be wrong: the number of data may be too small or the variability is too large to detect a possible difference. In both cases, it is useful to calculate the probability of rejecting the null hypothesis, stating a predefined difference between the groups and error variance structure.

Power analysis for MANOVA can be applied and can be performed using the Pillai-Bartlett's trace. It starts from calculating a hypothetical between- and within-group sum of squares matrix (see Appendix A5). The hypothetical between-group sum of squares matrix $\underset{\sim}{\mathrm{H}}_1$ is calculated as the crossproduct $\underset{\sim}{\mathrm{M}}^{\mathrm{T}}\underset{\sim}{\mathrm{M}}$ of the following matrix:

$$\underset{\sim}{\mathrm{M}} = \begin{bmatrix} \mu_{\mathrm{11}} & \mu_{\mathrm{21}} & \cdots & \mu_{\mathrm{p1}} \\ \mu_{\mathrm{11}} & \mu_{\mathrm{21}} & \cdots & \mu_{\mathrm{p1}} \\ \cdot\cdot & & \cdots & \\ \mu_{\mathrm{11}} & \mu_{\mathrm{21}} & \cdots & \mu_{\mathrm{p1}} \\ \mu_{\mathrm{12}} & \mu_{\mathrm{22}} & \cdots & \mu_{\mathrm{p2}} \\ & & \cdots & \\ \mu_{\mathrm{1r}} & \mu_{\mathrm{2r}} & \cdots & \mu_{\mathrm{pr}} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{\mathrm{1r}} & \mu_{\mathrm{2r}} & \cdots & \mu_{\mathrm{pr}} \end{bmatrix}$$

where μ_{ij} is the mean of group j for sample i (i=1,...,p; j=1,...,r), p is the number of samples and r is the number of groups. The hypothetical within-group sum of squares, or matrix $\underset{\sim}{\mathrm{W}}_1$, is the crossproduct of the residual values of the MANOVA model. If no correlation between measurements obtained by the same laboratory can be assumed, the matrix is a diagonal matrix, with the diagonal containing the expected variances of the parameter under interest for the different samples. Finally, let $\underset{\sim}{\mathrm{T}}_1$=$\underset{\sim}{\mathrm{H}}_1$+$\underset{\sim}{\mathrm{W}}_1$.

Then, the theoretical Pillai-Bartlett (PB) test is given by

$$\mathrm{PB}_1 = \mathrm{tr}(\underset{\sim}{\mathrm{H}}_1\underset{\sim}{\mathrm{T}}_1^{-1})$$

Next, the noncentrality parameter (λ) is given by [104, 109]

$$\lambda = \frac{\mathrm{PB_1}}{(\mathrm{s} - \mathrm{PB_1})/\mathrm{s}(\mathrm{n} - \mathrm{r} + \mathrm{s} - \mathrm{p})}$$

with s = min(p,r−1). If F_{crit} is the critical value of the F-distribution corresponding to the null hypothesis, for example the 95th percentile of the F distribution with p(r−1) degrees of freedom in the numerator and and s(n−r+s−p) numbers of degrees of freedom in the denominator , the power of the test is obtained by

$$\mathrm{Power} = \mathrm{P}\left\{\mathrm{F}\left[\mathrm{p}(\mathrm{r}-1), \mathrm{s}(\mathrm{n}-\mathrm{r}+\mathrm{s}-\mathrm{p}), \lambda\right] > \mathrm{F_{crit}}\right\}$$

Numerical example

Consider the determination of LDH in the Belgian EQA. There are 12 participants using Tris/EDTA following the Scandinavian, Italian and Dutch recommendations, of which 5 use a Beckman-Coulter kit and 7 use material from Olympus. Taking into account a minimal sample size of 6 for calculating z-scores, joining the two groups may not only enable us to evaluate the Beckman-Coulter users, but may also give more reliable estimates of analytical variability.

However, we only want to join the group if the results obtained by the two methods are sufficiently homogeneous. The formulas describing the influence of a shift on the standard deviation show that the standard deviation is increased by 10 % if the group means differ by 16 U/L. The influence of shifts of different size on the performance of z-score statistics and their ability to flag exceeding values (|Z|>3) is laid down in Table 2.1. Z-scores were calculated with the Grubbs' test based-approach as described in Chapter 1 on a simulation of 10000 normally distributed data of size 12, mean 400 and standard deviation 15, of which 5 values showed a shift of different size (16, 32 or 64). Afterwards one value was replaced by an outlier at μ_2+3σ, μ_2+5σ and μ_2+7σ. Assuming that we allow a maximal shift of 10 % in the standard deviation, how many samples would we need to reject the null hypothesis of no shift if the shift is smaller than 16 and the standard deviation is expected to be 15 U/L ? Let us work with 3 samples with a hypothetical mean of 400, 450 and 500 U/L. The hypothetical matrix of means then becomes:

Table 2.1 Percentage of flags for z-scores beyond 3 when a part of the data are shifted, in absence and presence of outliers at different distances.

Outlier	Shift (U/L)		
distance	16	32	64
μ_2+0	1.60 %	0.50 %	0 %
μ_2+3σ	13.8 %	4.60 %	0 %
μ_2+5σ	74.2 %	48.6 %	2.20 %
μ_2+7σ	99.1 %	93.5 %	29.4 %

$$\underset{\sim}{\mathrm{M}} = \begin{bmatrix} 400 & 450 & 500 \\ 400 & 450 & 500 \\ 400 & 450 & 500 \\ 400 & 450 & 500 \\ 400 & 450 & 500 \\ 400 & 450 & 500 \\ 400 & 450 & 500 \\ 415 & 465 & 515 \\ 415 & 465 & 515 \\ 415 & 465 & 515 \\ 415 & 465 & 515 \\ 415 & 465 & 515 \end{bmatrix}$$

and

$$\underset{\sim}{\mathrm{H}}_1 = \begin{bmatrix} 1981125 & 2224875 & 2468625 \\ 2224875 & 2498625 & 2772375 \\ 2468625 & 2772375 & 3076125 \end{bmatrix}.$$

Considering that there is no correlation between measurements obtained by the same laboratory and that all samples are analyzed with a standard deviation of 15, the within group sums of squares matrix writes

$$\underset{\sim}{\mathrm{W}}_1 = \begin{bmatrix} 2700 & 0 & 0 \\ 0 & 2700 & 0 \\ 0 & 0 & 2700 \end{bmatrix}$$

Finally, the hypothetical Pillai-Bartlett trace is expected to be 1.0055 and λ becomes 18.20. The probability that the Pillai-Bartlett test yields a significant F-value is then 0.78, which is acceptable. We selected data of 3 samples that were sent out in the second half of 2010 and beginning of 2011 (C/9945, C/9479 and C/10137). A Grubb's test for outliers didn't indicate any presence of outliers when the data were considered per method. The P-value of the MANOVA test was 0.0008, which was strongly significant. Hence, we conclude that there is a significant method-dependent bias between these two methods and the two groups cannot be joined.

2.4 Method comparison studies

The statistical methods described so far in this chapter are of interest when the presence of matrix effects is assumed or known. It may also be of interest to assess commutability of sample material before it is used in EQA rounds or for other assessments of standardization. The techniques used here are based on the techniques to compare two analytical methods. The first part of this section will first focus on the statistical approaches to compare the bias between two analytical methods, and continue with a description of the approaches described in the literature to assess commutability of sample material.

2.4.1 Assessing standardization of an analytical method

It is a common procedure to compare the bias between two methods using a regression model. Data are collected from an experiment in which a relatively large set of samples are split in two. One half of the aliquoted samples is analyzed with one method, one half of the aliquoted is analyzed with another method. Analyses are preferably performed in similar settings, i.e. in the same laboratory at the same time under similar conditions. One may also assess the comparability of results in an EQA setting, with a limited number of samples and with numerous aliquots sent to several laboratories. Graphical methods can be used to assess differences between samples [12, 4], but linear regression methods are preferred [180]. One can depart from a least-squares linear regression model (see Appendix A1), in which results obtained by one method are considered as the independent, explanatory or X-variable, and results obtained by the other method are considered as the dependent, or Y-variable.

However, the linear regression analysis is subject to a series of assumptions that should be met before its results can be interpreted.

- The residuals should be identically and independently distributed following a Normal distribution. This entails amongst others that the variability of the points along the regression line should be the same over the whole measurement range. Independence of points can be assured by experimental set up, for example by measuring all the aliquots in similar conditions, and excluding carryover effects from one sample to the other sample. Normality of errors can be assessed by a normal quantile plot. Heteroscedasticity, i.e. non-equality of residual variability along the regression line, can be dealt with by introducing weighing coefficients in the regression equation [152].

- The measurement of the independent variable is expected to be performed without any variability, i.e. there should be absolute certainty about the values obtained by the methods which values are set along the X-axis. This requirement can only be assumed in case of measurements obtained by a reference method, and not when a routine method is used. In the following sections we present a set of alternative approaches which can be used if one or more of the assumptions are not met.

Orthogonal regression

Orthogonal regression can be applied whenever uncertainty exists about the true value of the independent (X) variable. It yields parameter estimates of the regression line that minimize the orthogonal distance between the observed points and the regression line. The difference between least squares and orthogonal regression is shown in Figure 2.2.

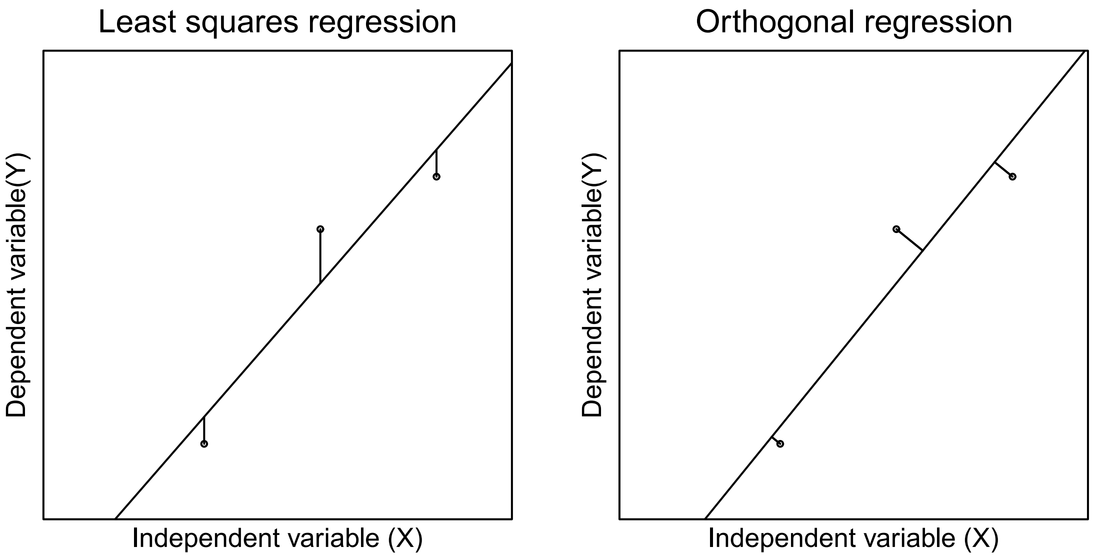

Figure 2.2 Least Squares and Orthogonal linear regression. The methods minimize a different distance between the observations and the regression line: Least Squares regression (left graph) minimizes the vertical distance, while orthogonal regression (right graph) minimizes the shortest distance.

There is no simple algebraic solution to the problem of orthogonal regression, although some solutions may be given if assumptions are made. It is often assumed that the variabilities of the independent variable and the dependent variable are known up to a fixed ratio. In the clinical chemistry field, this solution is often called Deming regression. It assumes that the values of the independent variable are normally distributed, and that the ratio of measurement uncertainties, expressed as variances, is known. Often, this ratio is set to 1, meaning that the analytical variability between the two methods is assumed to be the same. In its original form, Deming regression assumes that the measurement uncertainty is constant over the whole measurement range. Since this assumption cannot always be made, one can also consider general Deming regression [89].

Another approach is the standardized principal component [38]. It assumes that the ratio of the analytical error variance to the total variance is equal for the independent and dependent variabels.

Passing-Bablok regression

A method that doesn't require any special assumptions regarding the distribution of the data or the measurement errors of two methods was described by Passing and Bablok [9, 10]. It gives estimates for the regression line using an algorithm that is based based on the calculation of the slopes of the straight lines between any two points (See Appendix A3) . It also includes the calculation of confidence intervals, although it should be noted that they are usually wider than those obtained by classical linear regression [180, 9].

Stöckl [159] studied the question whether the statistical regression model or the analytical input data have more influence on the validity of the regression estimates. He found that the quality of the analytical input data is more crucial for the interpretation of the method comparison than the model used and recommends, in case of poor estimates from linear regression, to investigate the poor analytical performance rather than applying other regression models. He confirms that the sample concentrations should be adequately distributed over the whole measurement range and hence, the Deming, standardized principle components or Passing-Bablok regression are not always adequate for analyzing data obtained in EQA rounds, since mostly a limited number of samples is analyzed multiple times in this context.

A method for multiple analysis of a limited number of samples

Assessing method performance with respect to bias and variability can be done in an EQA setting by first determining a target value for a series of samples followed by sending aliquots to the participating laboratories for routine analysis. Since in this case only a limited number of samples are analyzed multiple times, Passing-Bablok or Deming regression in its different ways are not appropriate. We suggest a method based on the linear regression model and accommodated for possible violations of its basic assumptions, i.e. accommodating for a possible different measurement variability over the measured range and for a possible lack of linearity. The latter is mostly important in case some samples would exhibit matrix effects. The assumption of linearity may be performed by a lack of fit test [105]. The test compares a model where each group of measurements for a certain x_i is considered as a separate group, with a model where the different values x_i are considered to be continuous. For both models, the residual sums of squares (SSE) are calculated and used to obtain an F-statistic. It is a powerful test and may reject the null hypothesis solely based on statistical, and non clinical, reasons. For this reason, one may adopt it by allowing a minimal deviation from linearity for each sample. If we allow a deviation of δ % for each sample, we may bootstrap [34] the sample under the conditions of the null distribution, that allows a deviation of δ %, and proceed as follows:

(1) Obtain estimates of the weighted regression coefficients a and b

(2) Obtain the predicted values for the regression model $\widehat{\mathrm{y}}$

(3) Multiply each $\widehat{\mathrm{y}}$ with a randomly selected number from the interval $[1 - \delta/100;1 + \delta/100]$ and add a randomly generated value from the Normal distribution with 0 as mean and the residual standard error as standard deviation; As such we obtain n x m sets of data ($\mathrm{x_{ij}}$, $\mathrm{y^*_{ij}}$),

i=1,...,n; j=1,...,m with n the number of samples and m the number of laboratories

(4) Take a bootstrap sample from the n x m sets (x_{ij}, y^*_{ij})

(5) Calculate an F-statistic based on the bootstrapped sample

(6) Repeat "steps 2-4" 1000 times.

The distribution of 1000 F-values may serve as a sample of the null distribution while allowing a deviation of δ% for each sample. We conclude non-linearity if F calculated on the original data is larger than the (1-α)th quantile of the bootstrapped sample. Note that, in case of heteroscedasticity, the formulas can be easily rewritten taking into account weighing factors.

2.4.2 Assessing commutability of EQA samples

Several protocols have been proposed to assess commutability of sample material in quantitative terms. As for assessing standardization between methods, mathematical assessments are preferred above visual inspections via graphs. Vesper [180] has given an exhaustive overview of the applied methodology. In gross, the applied statistical methodology is a biplot, where samples can be visualized in clusters and where commutable material belongs to the same cluster as human sera, or a linear regression model. Eckfeldt et al. [33] used linear or polynomial regression to assess the commutability of materials used in EQA schemes. In this approach, simple linear regression analysis is performed to establish the relationship between results obtained from authentic patient samples using two measurement procedures and the two-tailed 95% prediction interval was calculated for the distribution of patient results. Measurement results obtained with a reference material (RM) are then compared against the 95% prediction interval. Linearity of the relation is assessed by including a polynomial terms of second order. When the latter reflects a non-significant effect, linearity is assumed. A power analysis could be performed by considering the prediction interval for a new point along the linear regression line [86].

An alternative approach based on evaluation of the residuals from regression analysis was introduced by Franzini to evaluate commutability. In this approach, a regression line is first calculated based on the two measurements of the clinical samples. Subsequently, the residual value for each couple of measurements of the RM with respect to the regression line of the clinical samples is calculated. Then, the residuals are standardized by dividing them by the residual standard error from the linear regression model of the clinical samples. Commutability is assumed if the absolute value of the standardized residuals is smaller than 3.

Another approach was suggested by Ricos et al. They used regression analysis with an evaluation approach based on expressing the residual as a percent to identify non-commutable control materials for creatinine. They used Passing-Bablok regression to determine the relationship between results for native patient samples. For each control material, a residual was determined as the difference from the value for that material predicted from the regression line for the native patient samples. The residual for the control material was expressed as a percent of the value predicted from the regression relationship and called a bias (in percent). The bias (in percent) for each control material was compared to three criteria to evaluate commutability [134, 135].

Baadenhuijsen et al. [8] described an alternative study design. The participating laboratories were grouped in couples of two and fresh patient sera were split into two portions, one portion from each sample was transported the same day to the partner laboratory, which in turn proceeded in the same way for its patient specimens. The interchanged fresh patient samples were then analyzed (within 24 h of the initial analysis) in the same analytical run with the reference materials, which were sent beforehand to each participant on dry ice. Relations between measurements of routine methods were assessed by a Passing-Bablok regression and results of sample material far away from the regression line between two laboratories were considered as non-commutable [8].

2.5 Application: Oestradiol and Progesterone

In the last decades, immunoassays have replaced the time consuming radioimmunoassays for determining steroids. They are based on chemiluminescence techniques and are executed on highly automated platforms, which are now routinely used to measure steroids, such as estradiol (E_2) and progesterone (P). The results of these methods, however, are not absolutely reliable, as demonstrated by a high imprecision and indeterminate accuracy. Boudou found for example that most of the immunoassays for progesterone showed an underestimation at high levels and an overestimation at low levels [17]. In addition, the kinetics of the antigen-antibody reaction and thus the measurement data may be influenced by matrix effects. Results of steroid immunoassays may also vary because of the procedure used to displace the substance from its serumprotein binding. In direct assays, displacement by competitive agents may be incomplete, whereas methods involving solvent extraction may suffer from variable extraction efficiency.

A study was performed at the Institute of Public Health of Belgium in 2005 to assess bias and variability of the routine clinical methods for Estradiol (E_2) and progesterone [27]. In an attempt to exclude matrix effects as much as possible, samples were off-the-clot serum and sent, without any additives or preservatives added, to the participants on dry ice. Five samples were obtained from normally cycling women and three were from a pool of donations by at least 10 different pregnant women. The sera from the normal cycling women were prepared with samples from one single donor; they were rapidly isolated to avoid hemolysis and kept in sterile conditions. The lipemia of the samples was evaluated by measuring the triglycerides on a Vitros system (Ortho-Clinical Diagnostics, Raritan, NJ, USA). The presence of other steroids was also analyzed: Estrone (Biosource Europe, Nivelles, Belgium), Estriol (Gamma, Angleur, Belgium) and 17a-hydroxyprogesterone (Biosource Europe). In order to control for possible bias caused by human anti-mouse antibodies (HAMA) in the two assays with the highest bias, two samples (E_2 concentrations of 1841 and 2026.4 pmol/L) were each measured in triplicate on Vitros and Vidas (Biomerieux, Marcy l'Etoile and Paris, France) systems before and after incubation with Heterophilic Blocking Tubes from Scantibodies Laboratory Inc. (Santee, CA, USA). Samples were stored at -70°C until the control samples were prepared for the trial. Sterile aliquots of 500 µl were prepared and stored at -70°C until they were distributed to the participating laboratories.

Determination of E_2 and progesterone reference value was performed by previously described Isotope Dilution - Gas Chromatography / Mass Spectrometry (ID–GC/MS) methods [173, 172, 174, 167]. The coefficient of variation (CV) of the target value of E_2, was estimated to be of the order <0.5% for concentrations >220 pmol/L and 0.6% for concentrations <220 pmol/L. For the extremely low concentrations (<18 pmol/L), the CV was about 2%. For progesterone, the CV of the target value is estimated at <1%. An overview of E_2 and progesterone concentrations, obtained by ID-GC/MS, with their clinical interpretation is given in Table 2.2.

The immunoassay systems taken into account were those most frequently used in Belgium: Advia Centaur (using assay ACS Centaur E2 6)(Bayer, Tarrytown, NY, USA), Immulite (DPC, Los Angeles, CA, USA), Elecsys (Roche, Basel, Switzerland), Access (Beckman-Coulter, Fullerton, CA, USA), Vitros and Vidas.

First, outliers were excluded for further analysis. One value deviating by more than 10000% from its true value, and the results of one laboratory of which more than half the results deviated by more than 50%, were disregarded. As in a clinical setting, where results are interpreted according to the patient's physiological status at the time of sampling, CV and bias are discussed per physiological condition with the emphasis on the highest CV or bias. A weighted linear regression analysis and lack-of-fit test were performed for each method separately to assess a linear relationship and the relation between concentration and bias [152]. The null distribution of the lack-of-fit test was developed by a 1000-fold bootstrap [34], which showed significance

Table 2.2 Overview of individual sample hormone concentrations as assessed by ID-GC/MS. These data are the target values of E_2 and progesterone, for comparison with the results obtained in the different laboratories.

Phase	E_2 (pmol/l)	Progesterone (nmol/L)
Early follicular	198.1	0.56
	209.4	0.80
	406.0	6.17
Perioovulatory or mid-luteal	598.1	22.53
	778.1	24.26
Pregnancy	1841	117.9
	2026	41.48
	3417	69.49

only if the means deviated by more than 10% from a linear relationship. To achieve satisfactory power, only results from those immunoassay methods that had at least seven users were considered in the study.

As a result, 140 laboratories were considered for E_2 and 155 for progesterone. For the eight samples in the study, none of the laboratories reported data below detection or quantification limit. Samples from different donors of which the concentration was defined by ID-GC/MS allowed a regression model to be used with the reference value as independent and the results reported by the laboratories as the dependent variable.

Three parameters were considered from the regression analysis: (i) P-value of a liberal lack-of-fit as a test for linear relationship: samples were omitted until a linear relationship was found, allowing for a deviation of 10 % for each sample; (ii) intercept of the weighted linear regression method with the hypothesis test whether there is significant difference from 0; and (iii) slope of the weighted linear regression method with the hypothesis test whether there is significant difference from the 45°-line.

Table 2.3 CV (%) per sample for E_2 for the six automated immunoassay analyzers most frequently used in Belgium in 2005.

Target value	Advia	DPC	Elecsys	Access	Vitros	Vidas
(pmol/L)	Centaur	Immulite				
198.1	24	21	11	23	24	15
209.4	24	14	11	49	22	16
406	11	12	7	10	16	7
598.1	14	11	7	18	11	7
778.1	22	11	8	12	13	12
1841	21	12	5	18	8	11
2026	29	10	5	10	15	22
3417	8	9	4	7	7	10

Table 2.4 Relative bias (%) per sample for E_2 for the six automated immunoassay analyzers most frequently used in Belgium in 2005.

Target value	Advia	DPC	Elecsys	Access	Vitros	Vidas
(pmol/L)	Centaur	Immulite				
198.1	7	-5	5	30	15	9
209.4	-12	-4	15	22	18	20
406	-4	-8	7	17	-11	4
598.1	9	-17	7	36	-26	0
778.1	14	-3	22	16	-12	10
1841	-4	-6	18	-10	2	43
2026	-13	25	6	20	96	239
3417	-4	-11	9	-6	-15	21

Table 2.5 CVs (%) per sample for progesterone for the six automated immunoassay analyzers most frequently used in Belgium in 2005.

Target value	Advia	DPC	Elecsys	Access	Vitros	Vidas
(nmol/L)	Centaur	Immulite				
0.56	107	59	52	202	88	102
0.80	58	43	23	84	33	74
6.17	16	11	6	33	9	10
22.53	8	10	7	18	9	12
24.26	8	8	7	11	7	9
41.48	16	8	11	15	9	10
69.49	8	7	7	10	7	10
117.9	14	14	8	45	10	6

Table 2.6 Relative bias (%) per sample for progesterone for the six automated immunoassay analyzers most frequently used in Belgium in 2005.

Target value	Advia	DPC	Elecsys	Access	Vitros	Vidas
(nmol/L)	Centaur	Immulite				
0.56	618	224	475	643	841	327
0.80	17	54	49	203	103	145
6.17	64	22	-23	81	-10	21
22.53	35	15	12	63	30	47
24.26	40	7	12	40	15	52
41.48	145	9	67	20	73	75
69.49	28	-8	16	7	8	33
117.9	28	15	34	33	35	46

Variability
The results are shown in Table 2.3 for E_2 and Table 2.5 for Progesterone. Considering the lack of precision for E_2 measurements of women in the early follicular phase, Elecsys, Vidas and Immulite had the lowest variability (with maximum values at, respectively, 11, 16 and 21%). Access had the highest variability, with CV up to 49%. CVs for Advia Centaur and Vitros were intermediate (24%). For progesterone, CV values and bias for concentrations <1 nmol/L were considered of no clinical importance and will not be discussed.

E_2 measurements of samples from women in the periovulatory or luteal phase demonstrated a similar trend. Elecsys, Immulite and Vidas had the lowest variability (respectively, 8, 12 and 12%), while Advia Centaur was the least precise (22%). Access and Vitros were intermediate with 18 and 16%, respectively. Progesterone immunoassays (Table 2.5) demonstrated maximum CVs <10% for Elecsys and Vitros (respectively, 7 and 9%). Access peaked at 33%, while the other systems showed intermediate CVs (Advia Centaur: 16%; Immulite: 11%; Vidas: 12%).

For the samples that were mixtures of single blood draws from pregnant women, Elecsys and Immulite had the lowest variability for E_2 (5 and 12%), while Vitros, Access and Vidas had, respectively, 15, 18 and 22%. Here, Advia Centaur had the highest variability (29%). For this category, CVs of progesterone were below or near 10% for Elecsys, Vitros and Vidas (respectively, 11, 10 and 10%). Notice that Access also has the highest CV (45%); Advia Centaur and Immulite are intermediate (16 and 14%).

Bias

The relative bias for E_2 for one sample (E_2 concentration 2026.4 pmol/L) was substantially higher than that seen for any other sample. For this sample, four methods showed a bias of 20% or higher: Access (20%), Immulite (25%), Vitros (96%) and Vidas (239%). Applying a theoretical correction for possible interfering substances (triglycerides, 2 mmol/L; estrone, 7013 pmol/L; estriol, 18 pmol/L) did not reduce the bias of the measurements to within acceptable limits (bias after correction by multiplying the cross reactivity coefficient with the concentration of interfering substances: Immulite: 25%, Access: 16%; Vitros: 78%; Vidas: 87%). Advia Centaur also showed high bias for progesterone (145%). Results of this sample were not included in the discussion of the bias. For E_2 (Tables 2.3 and 2.4), Immulite has negative bias for all other samples, while Elecsys and Vidas have overall positive bias. Concerning magnitude, Immulite had distinctively lower bias for the samples from women in early follicular phase (25%). Advia Centaur, Elecsys and Vitros had slightly higher bias (respectively, up to 212, 15 and 18%). Access and Vidas had up to 30 and 20%, respectively. For the samples from women in the periovulatory or luteal phase, Vidas performed best (bias <10%). Bias values for Advia Centaur and Immulite were slightly higher (respectively, 14 and -17%), while other systems have bias values >20% (Elecsys: 22%; Vitros: -26%; Access: 36%). The picture for samples from pregnant women is different: here, Vidas performed worst (43%). All other systems had intermediate bias (between 10 and 20%), and there was no clear difference between methods.

Bias values for progesterone are generally positive (Tables 2.5 and 2.6). Advia Centaur, Access and Vidas had overall positive bias. For all other methods there was only one sample with negative bias. All systems showed bias values

Table 2.7 Mean regression coefficients between reference values and values reported by the laboratories. Results are expressed in pmol/L for estradiol and nmol/L for progesterone.

Immunoassay	Estradiol				Progesterone			
	Intercept		Slope		Intercept		Slope	
Advia Centaur	17.82		0.96		2.37	*	1.26	*
DPC Immulite	-14.84	*	0.92	*	0.87	*	1.07	*
Elecsys	24.57	*	1.11	*	0.19	*	1.15	*
Access	104.0	*	0.93	*	3.56	*	1.15	*
Vitros	-13.67		0.92	*	1.47	*	1.10	*
Vidas	-30.16	*	1.31	*	0.89	*	1.41	*

*Intercept significantly different from 0; slope significantly different from 1

for all phases of at least 15%. Values peaked at 81% (Access), 64% (Advia) or 52% (Vidas).

The results of the linear regression are shown in Table 2.7. For E_2, only Advia Centaur and Elecsys had a linear relationship between reference and reported values over the whole range (Figure 2.3). For the other systems, the previously mentioned sample (concentration 2026.4 pmol/L) had significant bias and had to be omitted from the regression analysis to obtain a linear relationship.

Considering intercept and slope (Table 2.7), Immulite, Elecsys, Access and Vidas had an intercept that significantly differed from 0, indicating a concentration-independent bias especially important for the low concentrations. The slope differed significantly from 1 for all systems except the Advia Centaur. Slopes ranged from 0.92 (Immulite and Vitros) to 1.31 (Vidas). Here too, it should be noted that a non-significant difference for Advia Centaur may have been caused by the high uncertainty of the method. For progesterone (Figure 2.4), one system did not seem to suffer from a lack of linear relationship between RMV and reported values (Access). For the other systems, one (Advia Centaur, Elecsys, Vidas) or two (Immulite, Vitros) samples had to be omitted to obtain a linear relationship. Intercepts were all significantly higher than 0, ranging from 0.19 (Elecsys) to 3.56 (Access). Slopes were all significantly higher than 1, yielding deviations from 7% (Immulite) to 41% (Vidas).

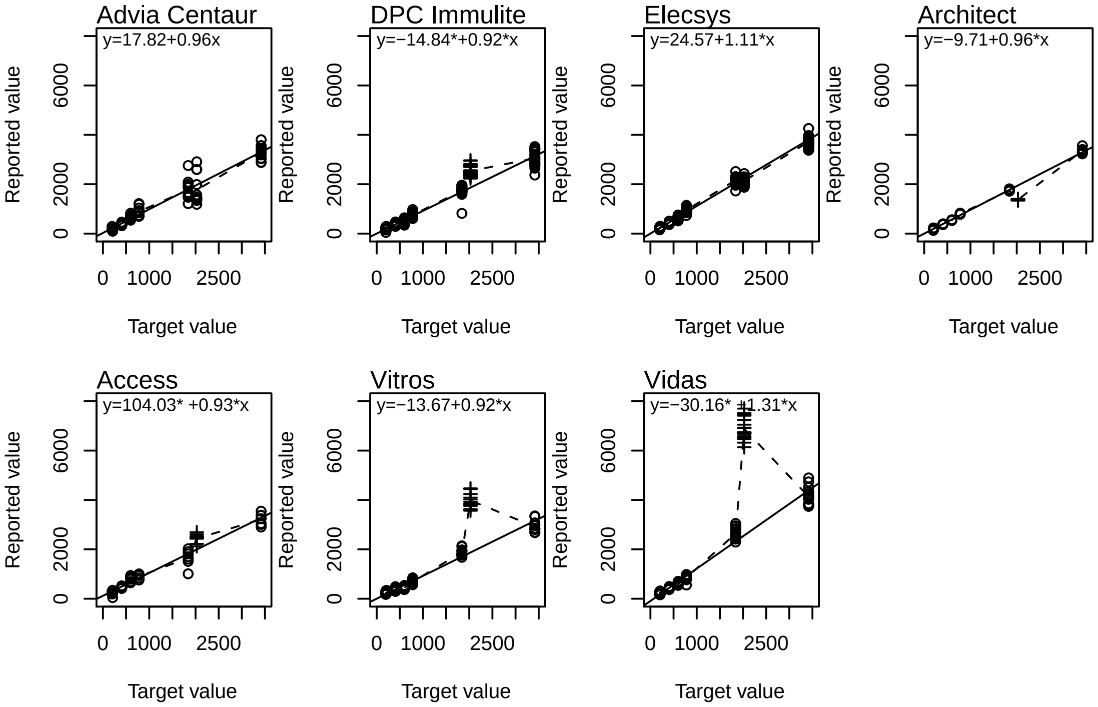

Figure 2.3 Weighted linear regression results for E_2. Values indicated by crosses (+) are not included for weighted linear regression, in order to obtain a linear relationship. All units are in pmol/L.

To conclude, we can say that mean bias values of 0.20% occurred in 19% of E_2 and in 90% of progesterone measurements. This suggests that the IVD directive from the European Union may not go far enough and that a more suitable requirement may be to strive to compare every method with the highest possible standard (in this case ID–GC/MS as reference method).

In addition, variation differed considerably between methods; it also differed considerably from reported intra-laboratory variance. Only for Elecsys were the reported intra-lab CVs [195, 11]) for both E_2 and progesterone comparable to the CVs in this study. This indicates that, for this system, the inter-laboratory contribution of variance to the total was very small. For

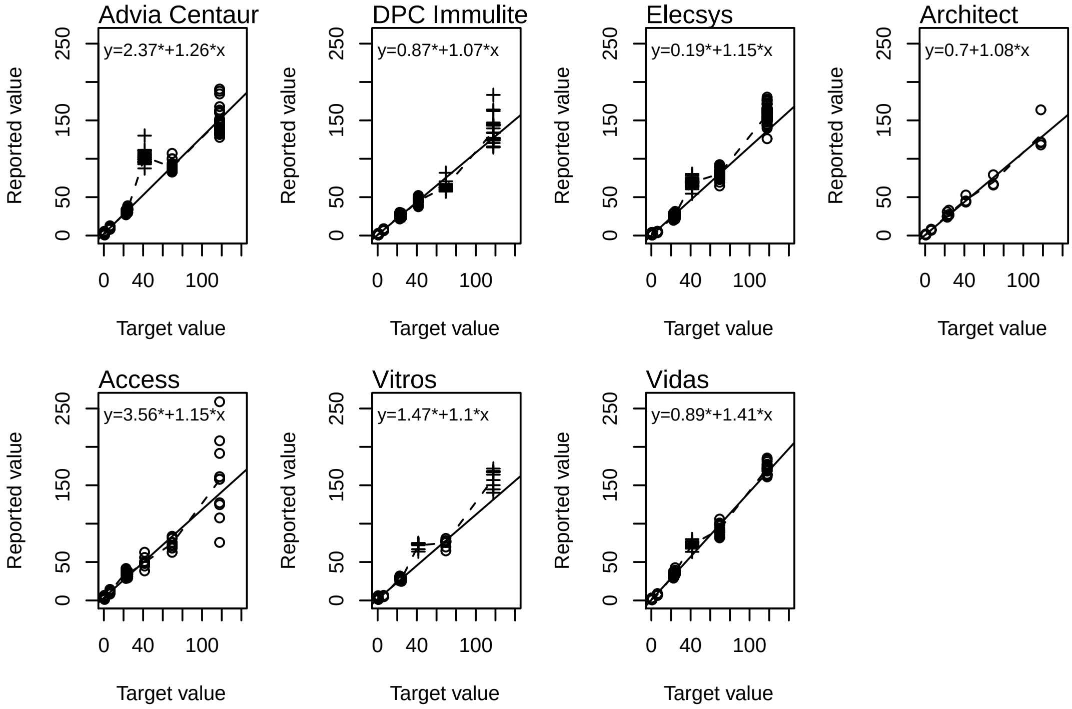

Figure 2.4 Weighted linear least-squares regression results for progesterone. Values indicated by crosses (+) were not included for weighted linear regression, in order to obtain a linear relationship. All units are in nmol/L.

other systems, the total inter-laboratory variability was clearly wider than the reported intra-lab uncertainties [138, 192, 51, 5, 171, 168, 90], which points to a significant difference between results for the same sample obtained in different laboratories.

The bias of several methods for particular samples jeopardized a linear relationship between RMV and routine methods for E_2 and progesterone. Vitros systems reported almost twice, and Vidas more than three times the concentration. Cross-reactivities reported by manufacturers were too low, however, to explain the behavior of the sample by interferences from another substance. Neither could lipemia or hemolysis be the reason, because the samples showed no tendency to behave oddly. A test in which possible HAMA [23, 72] were excluded before analysis did not help: Vidas still reported bias >200% and Vitros still showed >90% bias. To date we can offer no explanation for the abnormal bias seen for this sample with some methods. However, the sample consisted of a mixture of serum samples. Apart from the results recorded for that particular sample, it should be noted that no single method had bias values below 10% for all samples.

Considering progesterone, the results generally showed higher CVs and bias values. All methods had overall positive bias values, ranging to 40%. In conclusion, we can say that for E_2 and progesterone measurements, a linear relationship between the reference method values, determined by ID-GC/MS and reported values was not assured for most methods in a range for E_2 from 198.13 to 3417.3 pmol/L and for progesterone from 0.56 to 117.85 nmol/L. Overall precision for progesterone was better than for E_2 for all automated analyzers except for Access.

2.6 Conclusion

In an era of increased patient mobility, comparison between laboratory results of different centers, often obtained by different analytical methods, is of great importance and hence, standardization between methods should also be a priority. Interlaboratory comparisons, as carried out by EQA, have become essential to assess standardization between methods. There are however several practical aspects involved that make these kind of studies a tedious task.

EQA sample material should be made on a large scale and be treated for ensuring stability of parameters over time. Often, samples of different patients are merged and receive additives or other treatments to preserve them over a long time. These manipulations may induce matrix effects, i.e. the difference between analytical results obtained by different measurement procedures do not reflect the differences observed for routine clinical samples. As a consequence, samples may lack similarity with genuine samples routinely analyzed in the laboratory - an effect also called lack of commutabilty. Often a trade off has to be made between long-term stability of sample material and commutability.

For some parameters, commutable material exists, but unfortunatley, today, the only way to obtain material that is commutable for a large set of parameters is fresh, single-donation material. Freezing or pooling material from different patients may induce bias between methods. However, finding fresh serum single-donation samples is not easy, definitely when the EQA organizer wants to send samples that reflect a particular pathology or physiological stage. In addition, the intake of particular drugs may induce matrix effects as well. For this reason, EQA organizers should be aware of the degree of commutability of their sample material and undertake the necessary actions whenever commutability is not attained and laboratories should be partitioned into so-called peer groups. A peer group is a group of laboratories using the same or similar analytical methodology for the determination of a certain parameter, grouped in such a way that the group is free of matrix-dependent bias.

The homogeneity of the peer group with respect to matrix effects is a basic and necessary assumption that has to be assessed before any evaluation can be carried out, since a heterogeneity inflates the standard deviation of the peer group and may mask deviating results. Whenever details about the applied methodology are available and the EQA organizer can use results from previous surveys a multivariate model can be built that can detect heterogeneity with sufficient power. Before application of the model, the user has to set up a tolerance limit towards inflation of standard deviation. From this, a maximal bias between methods can be derived. The latter is subject to a power analysis, that yields the required number of samples used in the multivariate analysis to obtain with satisfactory probability the certainty that a peer group is homogeneous.

Another option to avoid matrix effects is a deliberate choice of sample material. We could think of a preferred hierarchy of sample material. The top of the hiearchy is best fitted for standardization studies: fresh samples sent out immediately after collection from the patient. Care has to be taken however that the patient is free of HAMA and is not taking any drugs that may cause interference with one or more analytical methods. Next are single-donation samples which have undergone freezing, followed by fresh pools of samples and freshly frozen pools of samples. At the bottom of the hierarchy are the samples which have been stabilized by additives and/or have undergone freeze-drying.

Matrix effects, however, are not the only obstruction in the quest for standardization between methods. In this context, it is also important that method comparison studies, outside of the EQA setting, are analysed with the right statistical techniques. Ordinary least squares regression suffers from different flaws and is not always the most appropriate technique. Orthogonal regression should be used, while the basic assumptions should be assessed as well. The statistical techniques used for analyzing EQA data can be divided in two groups: evaluations performed for each sample separately and evaluations taking into account different samples for which a target value has been set.

Whenever an analysis is desired taking into account the values reported for different samples, for which an assigned value has been calculated using a reference method, a linear regression analysis can be applied in order to assess the bias of a particular method over a concentration range. The nature of EQA data is structured to perform a lack of fit test, which could indicate absence of linearity for one or more samples and hence, the presence of matrix effects. Care should be taken however with the power of the test, since a statistically significant lack of fit may appear even when the difference of the mean reported values of a certain sample does not exceed limits of clinical relevance. For this reason, the null distribution of the lack of fit test should be set up using predefined limits of acceptability, where the bootstrap is a useful help. The application of this technique to a survey for oestradiol and progesterone has shown that pooled samples of pregnant women that were fresh frozen showed matrix effects. It was also evidenced that traceability towards a standard of higher order is not a satisfactory criterion for ensuring standardization between methods. Because of the considerable bias differences between methods, a clinical follow-up of a patient should always be performed using the same, fit-for-purpose and well-validated assay. Considering the large bias of some methods, it is recommended to use method-specific reference intervals for the different physiopathological conditions.

CHAPTER 3

Combined evaluation of different EQA surveys or samples

3.1 Introduction

Scientific articles reporting results from EQA have been focusing mainly on the evolution of the total variability between laboratories. Little attention has been paid to the individual follow-up of laboratories over a longer time period, or over a series of parameters. A literature search for scientific articles published in 2010 and in the first half of 2011 reporting findings of EQA surveys for at least one quantitative parameter for different samples resulted in 16 articles. Less than half (only 7 from the 16) mentioned an evaluation and follow up of the individual results of the laboratories, mainly by counting the number of laboratories reporting unacceptable results. All articles focussed in some or another way at the inter-laboratory variability. Although the latter is an important indicator of measurement uncertainty expressed on an inter-laboratory level, analyzing data over different surveys sheds light on the individual performance of the laboratories. A survey with one sample may inform the laboratory about its performance with respect to its peer group or to analytical performance goals. When an interpretation is made over several samples together, information can be generalized and a better understanding of the laboratory performance in terms of bias, variability and total error can be obtained, helping laboratories deciding where efforts for improvement should be made [35]. Only few approaches have been proposed in the literature to calculate performance statistics by combining data from different samples, surveys and/or laboratory test parameters. They will be presented, exemplified and commented in the next sections, with a focus on their robustness against outliers, and their ability to describe elevated bias and/or variability.

3.2 An artificial data set to compare the approaches

To evaluate the performance of the different approaches for finding poorly performing laboratories, they were applied to the same artificial data set. This data set represents simulated data from 224 laboratories, which have reported results for 10 parameters determined on one sample for 40 artificial surveys. It has been created to simulate the effect of outliers, variability and bias. For this purpose, a full factorial design was set up with the following factors and their levels:

(1) Outlier frequency: none (0%), low (5%) and high (10%)

(2) Bias: none, small (all data shifted 1 SD upwards), high (all data shifted 10 SDs upwards)

(3) Variability: none (SD=1), slightly increased (SD=2), highly increased (SD=5)

The data set contains results of 120 excellently performing laboratories, without outliers or bias and a standard deviation of 1. For each of the other combinations of the factor levels, 4 laboratories were simulated. Outliers, bias or variability were each time simulated for only one of the 10 parameters. This led to a data set of 224 laboratories, of which 120 performed excellently and 104 exhibited outliers, increased variability and/or higher bias for 1 parameter. Mean values of samples were arbitrarily set at a value between 10 and 70, yielding CVs between 1.4% and 45%. Note that about half of the laboratories exhibited a deviation from the ideal performance, i.e. their results contained outliers or reflected an increase in bias and/or variability. Although in reality a smaller fraction of the laboratories is expected to exhibit deviating results, this set up is preferred here to give a clear idea of the performance of the different evaluation methods with respect to weakly performing laboratories as explained by presence of spurious results, high variability or bias.

3.3 Variance and Bias index scores

3.3.1 Introduction

The principle of calculating a performance index on the basis of survey results has been adopted in the United Kingdom since several decades [188, 19] and is still a popular tool for expressing laboratory performance [118, 66, 170]. The cornerstone of the calculation is a comparison with a Chosen Coefficient of Variation (CCV; expressed as a percentage), in casu a mean Coefficient of Variation obtained for a certain parameter at a certain survey in the past. A "bias index score" (BIS) and a "variance index score" (VIS) is calculated by taking the difference between each participant's reported value (x) and a designated value (DV) and to express it as a fraction of the CCV, expressed as a percentage:

$$\text{Bias Index Score (BIS)} = \frac{(\text{x} - \text{DV})}{\text{DV}} \frac{10000}{\text{CCV}}$$

$$\text{Variance Index Score (VIS)} = |\text{BIS}|$$

The DV is calculated as a trimmed mean, whereas the BIS and VIS represent in fact the ratio, expressed as a percentage, of the relative distance between an individual result and its target value, with a coefficient of variation from the past. For excluding the effect of outliers, BIS values higher than 400 or lower than -400 are set at 400 and -400, respectively. Similarly, VIS values higher than 400 are set at 400.

Performance scoring is made using the VIS, which can be combined over surveys or analytes. Combining over surveys, the UK NEQAS defines a mean running VIS (MRVIS), which is the mean VIS value of the last 10 or 12 values. Consequently, a mean running BIS (MRBIS) is given as the mean BIS of the last 10 or 12 values. In addition, an overall overall mean running VIS (OMRVIS) is given as the last 40 (or 30) VIS values for all analytes assayed by the laboratory. Values of MRVIS or OMRVIS are categorized as shown in Table 3.1.

3.3.2 Example

The results of the MRVIS and OMRVIS are used to classify laboratories. In contrary with the UK NEQAS, the median is used as the DV, since this facilitates comparisons with other methods. As CCV value, the double of the CV of the population of non-contaminated result was proposed. Knowing that the expected mean of the absolute value of a standard normally distributed value is 0.80, the expected value of the MRVIS is then 40. The MRVIS and OMRVIS values were calculated for four distinct periods: the artificial survey 1-10, 11-20, 21-30 and 31-40. Subsequently they were classified according to Table 3.1 and the percentage of laboratories belonging to each class for the four different periods are shown in Table 3.2. Note that the distribution of

Table 3.1 Classification of laboratory performances according to their MRVIS or OMRVIS value.

Range of MRVIS, OMRVIS	Category
≤50	Ideal
50 - 100	Good
100 - 200	Adequate
>200	Poor

the number of laboratories over the different classes depends on the CCV value: a lower CCV value would have caused more laboratories belonging to the worse performance classes. In this case, it is more important to look at the relative shifts between the performance classes in presence of a combination of outliers, bias or increased variability, instead of the actual distribution between ideal, good, adequate or poor for a given shift, standard deviation and outlier frequency.

The presence of outliers causes a shift towards the weaker performance classes, mainly when the bias and standard deviation are small. However, the most determining effects on the classification are the bias and standard deviation, clearly exhibiting a shift towards the weaker classes with increasing bias and variability. It should be noted that a small increase in standard deviation, from 1 to 2, shows a larger effect than a small increase of the bias, from 0 to 1. A combined effect of bias and standard deviation shows a shift towards the worse performing classes that is larger than the shift when only bias or increased standard deviation exist. A larger bias or increased standard deviation results in 100% poorly performing laboratories.

Remark that a weaker performance of 1 of the 10 parameters doesn't influence the OMRVIS profoundly: outliers, bias or increased variability only influences the distribution between the ideally and well performing laboratories.

3.3.3 Comments

One of the major advantages of the variance index scoring system is a simple indication of laboratory performance, which can be merged over a variable number of surveys and parameters. Its use of a chosen coefficient of variation reflects the idea that, compared to the past, a good performing laboratory should respond closer to the designated value. In this way, the system favors laboratories performing better over time. Nevertheless, the evaluation strongly depends on the representativeness of the CCV. Since CCV values

Table 3.2 MRVIS and OMRVIS classifications for the example data set. MRVIS was calculated based on the parameter for which outlier frequency, variability and bias was induced, OMRVIS was calculated based on all the parameters. Results based on 224 simulated laboratories, having reported values for 10 parameters during 40 surveys.

			MRVIS				OMRVIS		
Bias	SD	Outliers	Ideal	Good	Adequate	Poor	Ideal	Good	Adequate/
		(%)							Poor
0	1	0	72.5	26.9	0.63	0.00	94.2	5.83	0.00
0	1	5	50.0	31.3	18.8	0.00	81.3	18.8	0.00
0	1	10	37.5	31.3	25	6.25	68.8	31.3	0.00
0	2	0	25.0	56.3	18.8	0.00	81.3	18.8	0.00
0	2	5	18.8	31.3	50	0.00	75	25.0	0.00
0	2	10	12.5	43.8	43.8	0.00	75	25.0	0.00
0	5	0	0.00	6.25	62.5	31.3	31.3	68.8	0.00
0	5	5	0.00	0.00	43.8	56.3	25	75.0	0.00
0	5	10	12.5	6.25	31.3	50	25	75.0	0.00
1	1	0	56.3	37.5	6.25	0.00	100	0.00	0.00
1	1	5	25.0	56.3	12.5	6.25	62.5	37.5	0.00
1	1	10	37.5	56.3	6.25	0.00	81.3	18.8	0.00
1	2	0	12.5	68.8	18.8	0.00	87.5	12.5	0.00
1	2	5	6.25	62.5	31.3	0.00	81.3	18.8	0.00
1	2	10	12.5	37.5	50.0	0.00	75.0	25.0	0.00
1	5	0	6.25	0.00	50.0	43.8	37.5	62.5	0.00
1	5	5	0.00	12.5	43.8	43.8	25.0	75.0	0.00
1	5	10	6.25	0.00	62.5	31.3	25.0	75.0	0.00
10	1	0	0.00	0.00	0.00	100	0.00	100	0.00
10	1	5	0.00	0.00	0.00	100	0.00	100	0.00
10	1	10	0.00	0.00	0.00	100	0.00	100	0.00
10	2	0	0.00	0.00	0.00	100	0.00	100	0.00
10	2	5	0.00	0.00	0.00	100	0.00	100	0.00
10	2	10	0.00	0.00	0.00	100	0.00	100	0.00
10	5	0	0.00	0.00	0.00	100	0.00	100	0.00
10	5	5	0.00	0.00	0.00	100	0.00	100	0.00
10	5	10	0.00	0.00	0.00	100	0.00	100	0.00

are a measure of the past, performance indicators denote only a reference to a previously attained analytical performance level.

Although only the VIS indicator is used to assess a laboratory's performance, the BIS and its derived variables are also given to the laboratory. It should be noted here that, although a combination of BIS with high absolute value and a low VIS reflects a bias for a certain parameter or laboratory, a high VIS value not solely means an increased variability. Disregarding outliers, the VIS and its derived parameters are a measure of the sum of bias and variability, i.e. total error. Hence, the term VARIANCE index score may be misleading. A laboratory experiencing an increased VIS should also look ad the evolution of the BIS values. An increased analytical variability can only be considered if the BIS doesn't increase. All other possibilities, i.e. an increased bias or a combination of increased bias and variability, cannot be distinguished. Although the method does an effort to make the VIS robust against outliers by ceiling individual VIS scores at 400, the method is still influenced by outliers and an observed increase in VIS or BIS may be due to an increased error rate in the laboratory as well. At last, pooling VIS values for different parameters together may mask laboratory mistakes in one single parameter, as seen by the high occurrence of ideal and good values of the OMRVIS. The latter demonstrates the necessity of an evaluation for each parameter separately.

3.4 Mean ranking scores

3.4.1 Introduction

Ehrmeyer and Laessig [35] abandoned the traditional approach to derive the criterion for acceptable results from a measure of variability based on the participant's results or obtained via analytical performance goals and developed a method that is related with non-parametric statistics. The basis of the test is the difference between an individual value and the "target" mean value of a parameter determined for each peer group:

$$\text{error} = \text{individual value} - \text{target value}$$

For each sample, laboratories are sorted from the smallest to the largest absolute value of the error and for each individual result, the proportion of results with equal or smaller absolute value of the error are recorded. As such, a cumulative distribution function is obtained and every reported value can be linked with its own percentile value, i.e. the percent of laboratories performing equal or worse with respect to the analysis of a certain parameter for a certain sample. Subsequently, median percentile values obtained for a certain laboratory for different samples are taken and a score on a scale from 0 to 100 is obtained. Lower values indicate good performance, higher values point to weak performance. By taking the percentile function result of a maximal allowable absolute error, comparison with analytical targets can easily be introduced into the algorithm and in this way, the approach is able to evaluate laboratories with respect to peer-group comparisons and to analytical target performance as well. Furthermore, averaging the errors may help laboratories differentiating between random and systematic bias: an average of the errors strongly deviating from zero indicates systematic bias, while a strongly deviating average of absolute values points to a combination of random error and/or systematic bias. Ehrmeyer has applied the method to EQA surveys for blood-gas and confirmed that the method can be applied to other data as well.

3.4.2 Example

The mean ranking approach has been applied to the example data set. Percentile functions were calculated for the reported values of the first parameter and median values were taken over 4 distinct groups of 10 surveys (surveys 1-10, 11-20, 21-30 and 31-40). Average ranking values of laboratories for each group of 10 surveys are given in Table 3.3. The effect of added outliers is almost not visible in the mean ranking. Only for the laboratories not having any bias or increased variability, a slight increase of the mean ranking can be observed when outlier frequencies rise. Further, ranking clearly increases with increasing bias or variability. In contrary to the performance index scoring system, a combined effect of bias and variability doesn't make the mean ranking to rise higher than their separate effects. For example, when outliers are not present, mean ranking of laboratories having a bias of one standard deviation and a standard deviation of 5 (mean ranking=60.9) is lower than when bias is absent (mean ranking=66.2). Furthermore, a high bias yields

Table 3.3 Mean ranking of laboratories exhibiting different deviations from the ideal process for the example data set, containing results of 224 simulated laboratories, having reported values for 10 parameters during 40 surveys, with induced bias, increased variability and outliers.

	Standard	Outliers	Mean
Bias	deviation	(%)	ranking
0	1	0	33.4
0	1	5	36.0
0	1	10	44.6
0	2	0	40.5
0	2	5	49.3
0	2	10	50.0
0	5	0	66.2
0	5	5	73.8
0	5	10	63.3
1	1	0	37.9
1	1	5	39.1
1	1	10	37.5
1	2	0	47.2
1	2	5	51.0
1	2	10	49.8
1	5	0	60.9
1	5	5	67.3
1	5	10	74.3
10	1	0	90.1
10	1	5	90.7
10	1	10	91.0
10	2	0	88.8
10	2	5	90.9
10	2	10	89.8
10	5	0	88.5
10	5	5	85.5
10	5	10	88.8

mean ranking values of around 90%, without distinction between variability levels or outlier frequencies.

3.4.3 Comments

A major advantage of the approach is its flexibility of use: it can be applied without the need to calculate any standard deviation or using an analytical performance goal. Even more, if analytical performance goals are available, they can be easily integrated. In addition, it is quite robust to outliers and hence, it defines weakly performing laboratories mainly on their performance in the analytical step of the total error process: an increase of the mean ranking value points to an increased bias and/or analytical variability. Laboratories having problems in their pre- or post-analytical phase, as may be expressed by a high frequency of outliers in EQA results, may escape notice by this test.The interpretation of the mean error values, before and after removing the sign, is very similar to the interpretation differences between BIS and VIS values from the performance scoring system and, as noted there, cannot always help in distinguishing between bias and variability.

3.5 Z-and u-scores count

3.5.1 Introduction

The Belgian EQAs for chemistry, immunoassays, therapeutic drug monitoring, alcohol, hematology and coagulation use a long-term evaluation technique based on the frequency of flags for z- and u-scores, as described by Albert [2]. They are calculated as shown in Table 3.4.

Table 3.4 Calculation and flagging of z- and u-scores.

	Calculation	Flagging
Z-score	$\frac{\text{individual result} - \text{group median}}{\text{group standard deviation}}$	$\lvert$z-score$\rvert>3$
U-score	$\frac{\text{individual result} - \text{group median}}{\text{group median}}$	$\lvert$u-score$\rvert>u^*$

u^*= analytical threshold, generally based on biological variability

Z-scores reflect the performance of a laboratory with respect to its peer group. Assuming that the majority of the laboratories reports a reliable result, a laboratory reporting a result that deviates more than 3 standard deviations away from the center of its peer group is flagged for a high z-score. The percentage of flagged u-scores reflects the performance of a laboratory with respect to a parameter-specific analytical quality specification limit, derived according to the Stockholm consensus and often based on biological variability (see Chapter 1).

Frequencies of flagged z- and u-scores (Pz and Pu) can be merged over the results reported by a laboratory for a certain domain, or over the results found with a certain method. Disregarding outliers, interpreting z- and u-scores together informs the laboratory about the quality of the applied methodology and about the quality of the use of the methodology, as outlined in Table 3.5. A combination of low z- and u-scores is optimal and points to the proper use of a proper methodology. When u-scores increase while z-scores are low, the laboratory is able to perform similar to other members of its peer group, so it uses its methodology in a proper way. The increasing u-scores indicate however an increasing failure to meet analytical targets, unveiling a weak method performance. When z-scores increase while u-scores are low, the laboratory is still able to meet analytical targets, but its results deviate from what other laboratories obtain, indicating a weak method use.

Table 3.5 Proper and improper use of methodology, ans explained by z- and u-scores.

	High u-scores	Low u-scores
High z-scores	Laboratory uses methodology improperly	Laboratory uses proper methodology improperly
Low z-scores	Laboratory uses improper methodology properly	Laboratory uses proper methodology properly

3.5.2 Example

The results are shown in Table 3.6. Z-scores were based on an average and standard deviation calculated after a removing outliers using Grubbs test ($\alpha = 0.05$), as described in Chapter 1. They were first calculated based on the parameter for which outliers, bias and variability were introduced and subsequently based on all the data. U-scores were evaluated using a maximal allowable deviation of 15 %. The percentage of z- flags and of an u-flags were averaged over each of the four groups of consecutive surveys (surveys 1-10, 11-20, 21-30 and 31-40). The percentage of flaggings is clearly subject to presence of outliers, as illustrated by an overall increasing percentage of citations when the frequency of outliers increases. The percentage of z-citations appears to be insensitive to laboratories experiencing increasing bias. On the other hand, laboratories experiencing an increase in their standard deviation clearly have a higher chance of being flagged for their z-scores.

The percentage of u-flags, however, clearly increases with increasing bias and variability and accumulates to higher percentages when bias and variability occur together. The difference between z- and u-flags for one parameter on the one hand and all the parameters on the other hand clearly demonstrates that the effect of outliers, bias or increased variability for one parameter becomes attenuated when more parameters are evaluated together.

3.5.3 Comments

The summary of percentage of z- and u-flags is an easy to understand and flexible approach to score laboratories. It can be merged over different samples, surveys, laboratories or methods and as such, it gives a good idea of the state of the art of the performance of a certain laboratory or method, the latter mainly with respect to the frequency of u-flags. They may indicated whether large deviations are due to a weak method, or a weak application of a good method. The statistics are however influenced by presence of outliers and cannot discriminate between bias or variability. The statistic can be used as a tool by EQA organizers to monitor the laboratories and to detect

Table 3.6 Percentage of z- and u-flags for the example data set, containing results of 224 simulated laboratories, having reported values for 10 parameters during 40 surveys, with induced bias, increased variability and outliers.

			One parameter		All parameters	
Bias	SD	Outliers (%)	Z-flags (%)	U-flags (%)	Z-flags (%)	U-flags (%)
0	1	0	0	1.08	0.25	0.200
0	1	5	7.50	7.50	0.75	0.813
0	1	10	10.6	11.0	1.1	1.06
0	2	0	0	5.00	0.43	0.563
0	2	5	6.25	12.5	1.00	1.38
0	2	10	6.88	17.5	0.75	1.75
0	5	0	0.625	36.3	0.18	3.63
0	5	5	5.63	41.3	0.75	4.13
0	5	10	13.1	43.8	1.8	4.44
1	1	0	0	1.25	0.19	0.188
1	1	5	6.87	8.13	0.81	0.813
1	1	10	7.50	9.38	0.94	1.00
1	2	0	0	6.25	0.44	0.625
1	2	5	3.75	13.8	0.63	1.38
1	2	10	11.9	16.9	1.50	1.88
1	5	0	3.13	39.4	0.44	3.94
1	5	5	6.88	41.3	0.88	4.19
1	5	10	11.3	45	1.3	4.56
10	1	0	1.25	99.4	0.44	10.1
10	1	5	8.13	99.4	1.2	10.1
10	1	10	13.1	100	1.6	10.2
10	2	0	2.50	95.0	0.63	9.69
10	2	5	11.2	96.9	1.4	9.69
10	2	10	11.2	98.8	1.3	9.93
10	5	0	26.4	81.3	2.8	8.18
10	5	5	18.4	86.3	1.9	8.769
10	5	10	24.4	85.0	2.6	8.56

laboratories suffering lack of performance and also here, evaluation per parameter is advised, since bad performance of one parameter may be masked when statistics are calculated over a whole range of parameters.

3.6 Long-term Analytical Coefficient of Variation

3.6.1 Introduction

A first evaluation technique based on a regression model was introduced by Meijer [93] in 2002 to evaluate the long-term performance of laboratories participating in the European Concerted Action on Thrombosis (ECAT) EQA programme. The model was initially applied to the test results for plasma antithrombin activity and later applied to the determination of activitity and antigen of protein C and S and the von Willebrand factor. [94, 95, 96]. The evaluation is done by fitting a linear regression model using consensus values obtained for a certain parameter for different samples as the independent variable (x) and reported values of the same laboratory as the dependent variable (y). See Appendix A1 for a detailed description of linear regression models. The slope (b) and the residual standard deviation (s) of each regression line were calculated, together with the mean values for x ($\overline{\mathrm{x}}$) and y ($\overline{\mathrm{y}}$) as well as the standard deviation of x ($\mathrm{s_x}$). The number of laboratory results included for a certain laboratory is expressed by $\mathrm{n_i}$.

First, the long-term total error (TE) for a laboratory i (i=1,...,N) is calculated taking into account all its reported values as follows:

$$\mathrm{TE_i} = \sqrt{\frac{\sum_{\mathrm{j=1}}^{\mathrm{n_i}} (\mathrm{y_{ij}} - \mathrm{x_j})^2}{\mathrm{n_i}}}$$

where $\mathrm{y_{ij}}$ is a reported value by laboratory i for consensus value $\mathrm{x_j}$(j=1,..,$\mathrm{n_i}$).

The total bias (B) and the random analytical error can be derived as follows:

$$\begin{aligned}\frac{\sum_{j=1}^{n_i}(\mathrm{y_{ij}}-\mathrm{x_j})^2}{n_i} &= \frac{1}{\mathrm{n_i}}\left[(\mathrm{b_i}-1)^2\sum_{\mathrm{j=1}}^{\mathrm{n_i}}(\mathrm{x_j}-\overline{\mathrm{x}})^2+\mathrm{n_i}(\overline{\mathrm{y_i}}-\overline{\mathrm{x}})^2+\sum_{\mathrm{j=1}}^{\mathrm{n_i}}(\mathrm{y_{ij}}-\hat{\mathrm{y}}_{\mathrm{ij}})^2\right]\\ &= (\mathrm{b_i}-1)^2.\frac{\mathrm{n_i}-1}{\mathrm{n_i}}\mathrm{s_x^2}+(\overline{\mathrm{y}}_\mathrm{i}-\overline{\mathrm{x}})^2+\frac{\mathrm{n_i}-2}{\mathrm{n_i}}\mathrm{s_i^2}\end{aligned}$$

where $\bar{\mathrm{x}}$ and $\mathrm{s_x}$ are respectively the mean and the standard deviation of the consensus values for the different samples, $\mathrm{s_i}$ the variability of the regression line, $\widehat{\mathrm{y}}_{\mathrm{ij}}$ is the value of the dependent variable of laboratory i for consensus value $\mathrm{x_j}$ as predicted by the regression model and $\mathrm{b_i}$ is the slope for the regression line. The formula contains three terms, of which only the last one is dependent on the variability of the points around the regression line. The two first terms make up the bias, while the latter term is an indicator of the within-laboratory variability. The factors $\frac{(\mathrm{n_i}-1)}{\mathrm{n_i}}$ and $\frac{(\mathrm{n_i}-2)}{\mathrm{n_i}}$ in the formulas are attributable to the difference in corrections for degrees of freedom in the definition of TE, s and $\mathrm{s_x}$. They approach 1 when n becomes larger and then the total error may be approximated by the following formula:

$$\mathrm{TE_i}=\sqrt{\mathrm{s_i^2}+(\mathrm{b_i}-1)^2\mathrm{s_x^2}+(\overline{\mathrm{y}}_\mathrm{i}-\overline{\mathrm{x}})^2}$$

The part of the total error attributable to the bias is:

$$\mathrm{B_i}=\sqrt{\frac{\mathrm{n_i}-1}{\mathrm{n_i}}(\mathrm{b}-1)^2\mathrm{s_x^2}+(\overline{\mathrm{y}}_\mathrm{i}-\overline{\mathrm{x}})^2}$$

The bias consists of a constant and a proportional, or concentration-dependent part. Constant (CB) and Proportional Bias (PrB) can then be written as:

$$\begin{aligned}\mathrm{CB_i} &= \sqrt{(\overline{\mathrm{y}}_\mathrm{i}-\overline{\mathrm{x}})^2}\\ \mathrm{PrB_i} &= \sqrt{\frac{\mathrm{n_i}-1}{\mathrm{n_i}}(\mathrm{b_i}-1)^2\mathrm{s_x^2}}\end{aligned}$$

The Long-Term Analytical CV ($\mathrm{LCV_a}$) is based on $\mathrm{s_i}$ and the mean value of all consensus values ($\overline{x}$). To allow comparison of the $\mathrm{LCV_a}$ among laboratories, it should be calculated after adjustment for the bias. Therefore, the $\mathrm{LCV_a}$ for a particular laboratory is now calculated using the formula:

$$\mathrm{LCV_a} = \frac{\frac{\mathrm{s_i}}{\mathrm{b_i}}}{\overline{\mathrm{x}}}.100\%$$

At last, the Analytical Critical Difference (ACD) reflects the minimum analytical capability of a laboratory to significantly distinguish between two different test results with a significance level of 95%. It is calculated as follows:

$$\mathrm{ACD_i} = \frac{\mathrm{s_i}}{\mathrm{b}}\sqrt{2\mathrm{Q_t}(0.975; \mathrm{n}-2)}$$

where $\mathrm{Q_t}(0.975; \mathrm{n}-2)$ is the 97.5th percentile of the Student's t-distribution on n−2 degrees of freedom.

3.6.2 Example

Results of $\mathrm{LCV_a}$, total, proportional and constant bias, total error and ACD are shown in Table 3.7. They were calculated and averaged over the four distinct consecutive survey periods as the other long term evaluation techniques (surveys 1-10, 11-20, 21-30, 31-40). The most striking results are the negative values for LCVa and ACD, caused by a negative slope of the regression line. A reason for a negative slope is visualized in Figure 3.1. The data set consists of 10 points, of which 9 follow a line close to the 45°-line. One point of the 10 is an extreme outlier, it is 10 times its true value. It influences the regression variability and the slope from the regression line considerably, making the residual standard error to increase and the regression coefficient to be negative.

Due to the heavy influence of outliers on the regression model, it is preferred to interpret $\mathrm{LCV_a}$ and ACD only for the cases where no outliers were added. As LCVa is more a measure of variability, it is clearly influenced by the

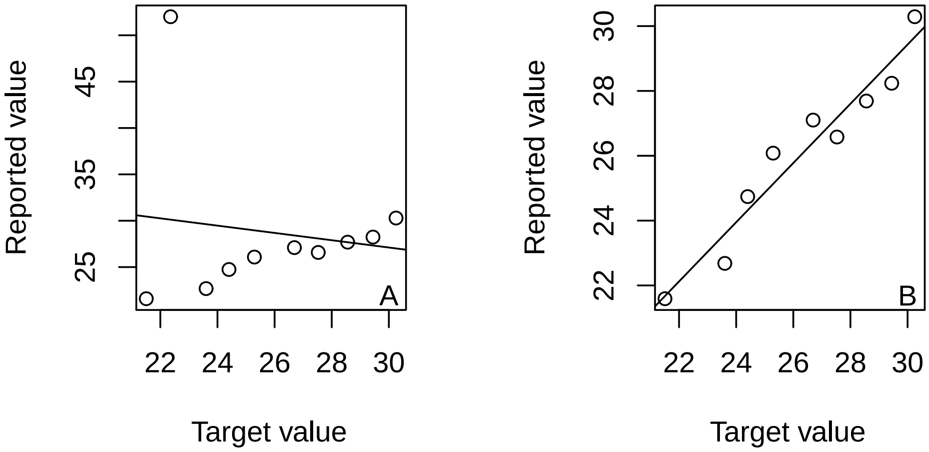

Figure 3.1 Regression line between target values and responded values, in the presence and absence of an outlier. Laboratory (A) has an outlier and an LCVa of 88.3%, laboratory (B) has reported the same values except the outlier, and has an LCVa of 3.1%.

analytical variability of the laboratory rather than the bias. We see however, certainly for laboratories having high analytical variability, the LCVa increasing with increasing bias. In terms of bias, a clear increase of bias is visible for increasing shifts. Remark that the constant bias follows the shifts well and that the proportional bias also increases with increasing analytical variability. The total error follows clearly the changes in bias and standard deviation and approaches to the square root of the sums of the squared shift and standard deviation. The ACD is heavily influenced by an increasing variability, but almost not at all by an increasing bias.

3.6.3 Comments

In comparison with the three previous approaches, this approach is able to estimate individual, within-laboratory bias and variability and doesn't take into account the values of other laboratories. The approach could even be used in small-scale surveys, because the performance indicators are calculated with respect to the reference values and not the values reported by other laboratories. To monitor the performance of a laboratory, one may interpret the values of total error (TE), bias, $\mathrm{LCV_a}$ and ACD with results previously obtained by the laboratory, results obtained by other laboratories

Table 3.7 Total Error (TE), Constant (CB) and Proportional Bias (PrB), LCVa and ACD for the example data set, containing results of 224 simulated laboratories, having reported values for 10 parameters during 40 surveys, with induced bias, increased variability and outliers.

Bias	SD	Outliers (%)	LCV_a	CB	PrB	TB	TE	ACD
0	1	0	3.70	0.503	0.25	0.601	1.09	3.35
0	1	5	-2.33	2.24	1.56	2.96	6.1	0.250
0	1	10	1.28	2.96	2.46	4.05	8.53	19.3
0	2	0	7.04	0.592	0.41	0.792	1.84	7.18
0	2	5	-11.5	1.94	2.28	3.10	5.85	-28.5
0	2	10	13.3	1.94	1.6	2.68	6.92	54.6
0	5	0	13.5	0.998	1.16	1.69	4.43	19.0
0	5	5	38.7	2.69	1.78	3.58	7.47	78.3
0	5	10	99.3	3.48	2.15	4.43	10.6	322
1	1	0	3.84	0.53	0.29	0.644	1.09	3.31
1	1	5	-30.7	2.54	1.77	3.21	6.58	-172
1	1	10	10.9	2.74	1.56	3.29	7.05	46.5
1	2	0	7.66	0.663	0.48	0.887	2.07	7.92
1	2	5	48.9	1.99	1.37	2.49	4.58	96.7
1	2	10	34.3	4.13	1.92	4.68	9.14	152
1	5	0	15.7	1.65	1.37	2.39	5.15	16.1
1	5	5	35.5	3.14	1.59	3.80	8.09	105
1	5	10	-42.0	3.18	3.42	4.99	10.1	-70
10	1	0	3.66	9.71	0.29	9.72	9.76	3.42
10	1	5	16.4	11.2	0.98	11.3	12.1	31.9
10	1	10	20.1	12.2	1.78	12.4	13.6	59.1
10	2	0	8.14	9.50	0.55	9.52	9.69	8.1
10	2	5	57.3	11.1	1.5	11.2	12.1	131
10	2	10	14.8	11.2	1.04	11.3	12.1	35.9
10	5	0	-20.3	9.99	1.24	10.1	11.1	-48.1
10	5	5	10.9	10.2	1.86	10.5	11.9	22.8
10	5	10	1.18	11.1	1.31	11.3	12.8	35.6

or by performance goals.

Foremost among the properties of the LCV_a and ACD is the high dependence of absence of outliers before a worthwile interpretation can be made. Outliers influence the estimates of the residual error and the regression line coefficients of a simple linear regression line heavily. Besides, a slight bias may influence the LCV_a and ACD in a non-intuitive way. The regression line of the results from a laboratory having a slight positive bias for the higher values and a slight negative bias for the lower values for example, will have a higher slope and hence a lower LCV_a than a laboratory having no bias but a comparable residual error.

Bias has been introduced in the data set as a constant bias and this is nicely reflected by the constant bias measurement of the method. However, estimations of proportional or constant bias are disturbed by high a high residual error, as is seen by an increasing total, proportional and constant bias with increasing variability. LCV_a on its turn is, disregarding outliers, a good indicator of increased variability. Bias doesn't influence LCV_a much. Although the technique easily allows combining information from different samples and surveys together, it lacks flexibility on the level of combining information obtained for different parameters together and hence, cannot be applied for a global evaluation of a series of parameters. It does however enable EQA organizers to monitor laboratory performance over a long period.

3.7 A novel three-step method

3.7.1 Introduction

The methods discussed so far show one or more of the shortcomings listed below.

Robustness against outliers
Most of the performance indicators suffer from lack of robustness against outliers. The LCVa is the most prominent example of how a single deviating point may make a performance indicator unreliable. In addition, the performance score indexes and the z-and u-scores counts score laboratories worse when outliers occur.

Distinguishing between outliers, bias and variability

Because of the pedagogic aspect of several EQA prgrammes, they should strive to help weakly performing laboratories to understand the nature of their mistakes. In this sense, a classification between different types of mistakes is preferred. Westgard [185] has pointed to the different character of mistakes in the pre- and post-analytical phase on the one hand and the analytical phase on the other hand. The first can more easily be represented by outliers occurring at a low frequency, while mistakes in the analytical phase can more easily be represented by increased bias and/or variability. Evaluation methods should strive towards a distinction between these three types of estimates.

Estimating within-laboratory variability

In EQA programmes, a measure of variability is often taken over a series of reported results for a certain sample, yielding a measure of total variability measure. The use and interpretation of this variability has two main drawbacks. First of all, the terminology may be misleading, since this measure of total variability is often called inter-laboratory variability. Following the rule of combined standard deviation as explained in the Guide to the expression of uncertainty in measurement (GUM) [194], the total variability found in data reported in EQA studies can be seen as a variability which is composed of an intra-laboratory and an inter-laboratory part. Confusion of the latter with the total variability should be avoided.

Secondly, the total measure of variability doesn't inform the individual laboratory about its proper analytical variability. A measure of variability reflecting the individual analytical variability, the intra-laboratory component, is needed. The latter may be calculated in two different ways. First, an EQA organizer may send two or more vials with the same content but a different label to the participants. After removal of outliers, a linear mixed model may be applied to the data, with the laboratories modeled as a random factor. Estimators of intra- and inter-laboratory variance can be obtained and their sum gives an estimate of total variance, which is a better estimator of the total variability than a variance of the reported values. Secondly, the EQA organizer may send different samples to the participants and determine a target value for each sample. A regression line can be calculated for each individual laboratory between the target values and the reported values. The residual error of the regression line is a measure of the analytical variability of the laboratory.

3.7.2 Pitfalls of using a linear regression line

In a first instance, a linear regression model made up between the target values of a series of samples and the data reported by a particular laboratory can be considered as a valuable model to estimate analytical variability and bias of a certain laboratory. Fig 3.2 demonstrates the model. The spread of the points around the regression line is measured by the residual error and reflects the analytical variability of the laboratory. The position of the regression line, indicated by its intercept and slope, with respect to the standard 45°-line, reflects the bias of the regression line. Bias is then expressed as the deviation between the intercept and slope of the individual regression line and the intercept and slope of the reference regression line, such as the mean regression line of a group of laboratories, or the 45°-line.

Remark that the linear regression model doesn't take into account the measurement uncertainty around the target values. As a consequence, target values should be determined with the highest accuracy. They can be set by a reference laboratory or be calculated using the reported values. We refer to Duewer [32] for an exhaustive overview of statistical estimators of target values in an EQA setting. To our experience, the median is a good estimator of central location and will be used in the examples given in this work. In addition, the regression model is only valid when matrix effects can be excluded and hence a linear relation between the target value and the responded values can be attained. For a detailed discussion about matrix effects, see Chapter 2.

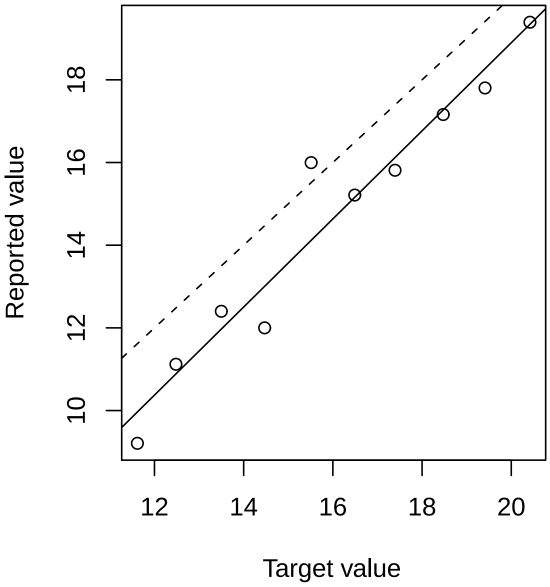

Figure 3.2 Example of a linear regression model. The independent variable is made up by the "Target values" of different samples, the dependent variable is set up by the "Reported values" of a particular laboratory. The solid line is the least-squares linear regression line through the points, the dashed line is the 45°-line.

The regression model is, as already illustrated by the LCVa, strongly influenced by outlying points. Figure 3.3 shows the relation between target and reported values for two different laboratories for the same group of surveys, as found in the example data set. The regression lines represented by the full lines in both graphs are drawn through all the points in the graph. They have the same residual error and regression coefficients. There is however a difference: while the majority of the points in graph A lie closely around a regression line (the dashed line) and only one point deviates strongly (the point indicated by the arrow), the majority of the points in graph B deviate strongly from the regression line. Recalculating the regression line in graph A without the exceeding point makes the residual error to drop from 42 to 2. The point indicated by the arrow in graph A is called an outlier, since it is a non-representative point, influencing the regression's coefficients and residual error. Because outliers around a regression line may exist, they should first be excluded before interpreting the regression line.

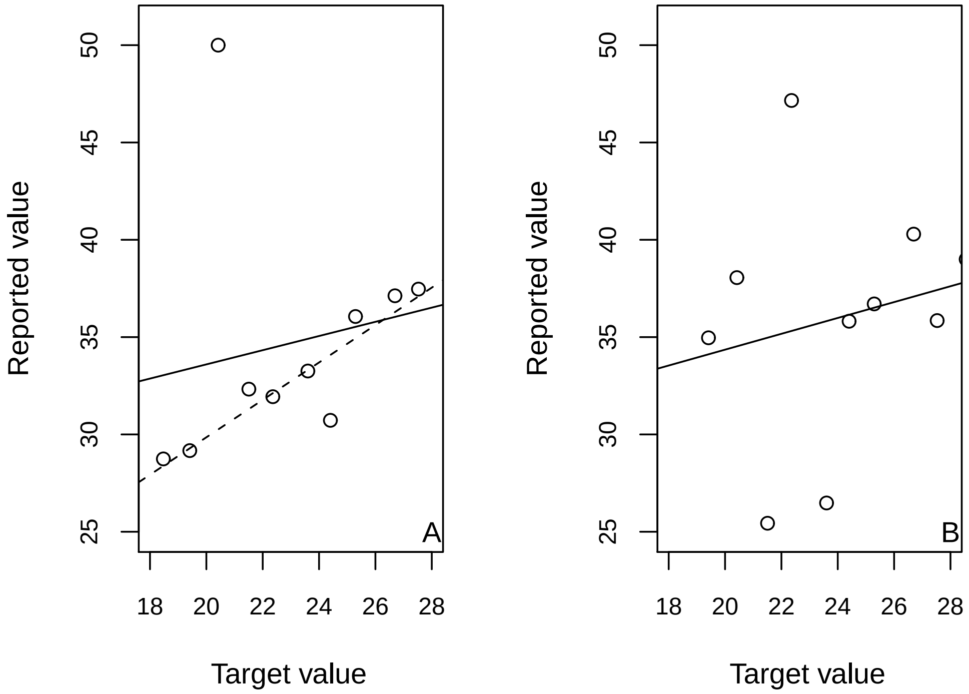

Figure 3.3 Effect of an outlying point on a least-squares linear regression line. The full lines are drawn through all the points, the dashed line in graph A is drawn without the exceeding point. The regression residual variability and position of the regression line for laboratory A and B are equal.

Two different types of solutions exist to deal with outlying observations when calculating a regression line. The first option calculates robust estimates of the regression line and its residual error, the second option searches for outliers and removes them before calculating a simple linear regression line. Atkinson [7] has shown that the second option is preferred, since robust estimates show a low estimation efficiency. Therefore, before any variability of bias is calculated by means of a least squares regression line, outliers against the individual regression lines have to be removed.

Once a regression line has been obtained after regression outliers have been excluded, one may interpret its residual error as a measure of analytical, within-laboratory variability and the intercept and slope as a measure of bias. Remark that there is a high uncertainty about the regression lines with high analytical variability, as depicted in Figure 3.4. Both regression lines have the same position, as indicated by their intercept and slope, but the certainty about the position of the latter is very low, i.e. there is a high probability that its position may be strongly different if results from other samples would have been used to draw it. For this reason, variability should be interpreted before interpreting any bias and an algorithm consisting of three steps can be built: first, outliers are found with respect to the individual regression line of each laboratory, followed by a step to find regression lines with excessive variability and finally regression lines with high bias are identified.

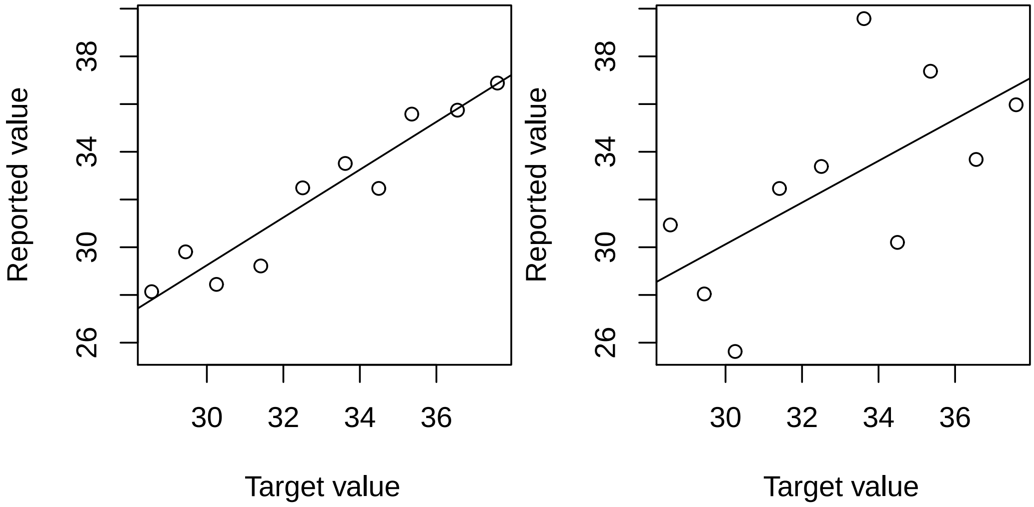

Figure 3.4 Regression lines with equal regression coefficients but different residual error. The position of the regression line in the left graph is more certain than in the right graph.

3.7.3 Description of the algorithm

The 3-step procedure [28] is based on a linear regression model. Let N denote the number of laboratories participating in the EQA programme and n_i the number of EQA samples assayed by laboratory i (i=1,...,N). For each laboratory, the relationship between the reported values and the target values of a given parameter can be described by the linear model

$$y_{ij} = a_i + b_i x_j + e_{ij}$$

where y_{ij} is the value reported by laboratory i (i = 1, ..., N) for sample j (j =1, ..., n_i), a_i and b_i the intercept and slope of the regression line for laboratory i, x_j the target value for sample j and e_{ij} the residual or error term. It is assumed to be normally and independently distributed with a mean of 0 and variance s_i^2. The latter is a measure of the analytical variability for laboratory i, while the parameters a_i and b_i indicate the laboratory's bias from the 45°-line, or from the mean regression line estimated from all laboratories.

Step 1. Finding outliers against the individual regression lines

The first step considers the regression line of each laboratory individually and searches for points that are exceedingly far from the regression line and can be considered as outliers. They are seen as measurements that are not representative for the analytical process of the laboratory. To exclude potential outliers, we propose to use a least trimmed squares (LTS) regression [141, 143] to obtain a rough estimate of the regression line, which holds for the majority of the data points. This is followed by use of the outlier search algorithm described by Atkinson [7]. LTS-regression minimizes the sum of the squared residuals of a predefined fraction of the data, usually a proportion slightly above 50%. The latter can be parametrized in software for calculating LTS regression. However, default values do not always yield an optimal curve and it is advisably to increase the fraction for which the sum of the squared residuals is minimized, trying to configure the proportion high enough for an optimal curve, but not too high in order to exclude outliers with high probability. One may proceed as follows:

(1) Define Q as the largest integer smaller than 60% of the number of couples (x_i,y_i) used to calculate the regression line.

(2) calculate an LTS-regression that minimizes the sum of the squared residuals of Q data points. Define s_Q^2 as the sum its squared residuals.

(3) Calculate an LTS-regression that minimizes the sum of the squared residuals of Q+1 data points. Define s^2_{Q+1} as the sum its squared residuals. Repeat the same for an LTS-regression that minimizes the sum of the squared residuals of Q+2 and Q+3 data points. Define s^2_{Q+2} and s^2_{Q+3} as their respective sum of squared residuals.

(4) If $\frac{s^2_{Q+1}}{s^2_Q}>10$, $\frac{s^2_{Q+1}}{s^2_{Q+2}}>10$ or $\frac{s^2_{Q+1}}{s^2_{Q+3}}>10$, choose the regression line that corresponds with the lowest value of s^2_Q, s^2_{Q+2} and s^2_{Q+3}.
Next, the m_i ($m_i \leq n_i$) points

Next, the m_i ($m_i \leq n_i$) points satisfying the following inequalit

$$\left|\frac{e_{ij}}{s^*_i}\right| \leq Q_t(1-\alpha_1; n_i - 2)$$

were subject to an ordinary least-squares (OLS) linear regression. In the above inequality, $Q_t(1-\alpha_1; n_i-2)$ is the upper α_1-quantile of the Student t-distribution with (n_i–2) degrees of freedom,

$$s^*_i = 1.4826\left(1 + \frac{5}{n_i - 2}\right)\sqrt{\text{median } e^2_{ij}}$$

and α_1 is a predefined significance level. Let r_{ij} be the residuals of the OLS regression and s_i its residual error. Then, outliers were defined as values for which

$$\begin{aligned} \frac{r_{ij}}{s_i\sqrt{1-h_{ij}}} &> t^* \text{ for the points included in the OLS regression} \\ \frac{r_{ij}}{s_i\sqrt{1+h_{ij}}} &> t^* \text{ for the points not included in the OLS regression} \end{aligned}$$

The critical value t^* is the upper-α_2 percentile from the Student t-distribution with $(m_i - 2)$ degrees of freedom and h_{ij} is the corresponding diagonal element of the hat matrix from the LTS-regression. The h_{ij}-values are the so-called leverages, which measure the influence of the individual point on the position of the regression line.

It should be noted here that this test does not include any inter-laboratory comparison. Outliers are identified per laboratory by comparing its reported results with each other and with the target values.

Step 2: Finding laboratories with high analytical variability

The differences between the different residual variances should be taken into account when assessing bias with the peer group's mean regression line. When outliers are excluded in the first step, a regression line is calculated for each laboratory separately and an estimate of the residual variance s_i^2 is obtained for each laboratory. The second step of the procedure is to identify laboratories with excessive s_i^2 values. The method used has been described earlier [3, 49]. It is based on the assumptions that both the theoretical residual variance s_i^2 and the estimated residual variance s_i^2 are log-normally distributed. The following formula can be derived:

$$\log \sigma^2_{\alpha_3} = \log \mathrm{M} - 0.5\log\left[1 + \tfrac{\mathrm{V} - 2\mathrm{kM}^2/(2\mathrm{k}+1)}{\mathrm{M}}\right]$$

$$+\mathrm{Q_z}(1-\alpha_3)\sqrt{\mathrm{V} - 2\mathrm{kM}^2/(2\mathrm{k}+1)}$$

where $\mathrm{k}=\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{m_i}-1)^{-1}/\mathrm{N}$, M=Mean s_i^2, V=Var s_i^2 and $\mathrm{Q_z}(1-\alpha_3)$ is the upper α_3-percentile of the standard Normal distribution. Mean s_i^2 and Var s_i^2 are the mean and variance of the residual variances from the regression lines obtained in the first step. The value $\sigma^2_{\alpha_3}$will be used as the critical threshold above which a residual variance s_i^2 will be considered to be exceedingly large. Since some of the N estimates s_i^2 may themselves be outliers, a trimming procedure is performed before calculation by disregarding an arbitrary proportion of the lowest and highest variance values.

Step 3: Finding laboratories with exceeding bias

As before, let a_i and b_i be the estimated intercept and slope of the corresponding regression lines once outliers are excluded in the two previous steps. The literature on regression explains how the joint distribution of intercept and slope can be considered as bivariate Normal [152], with mean (a, b),

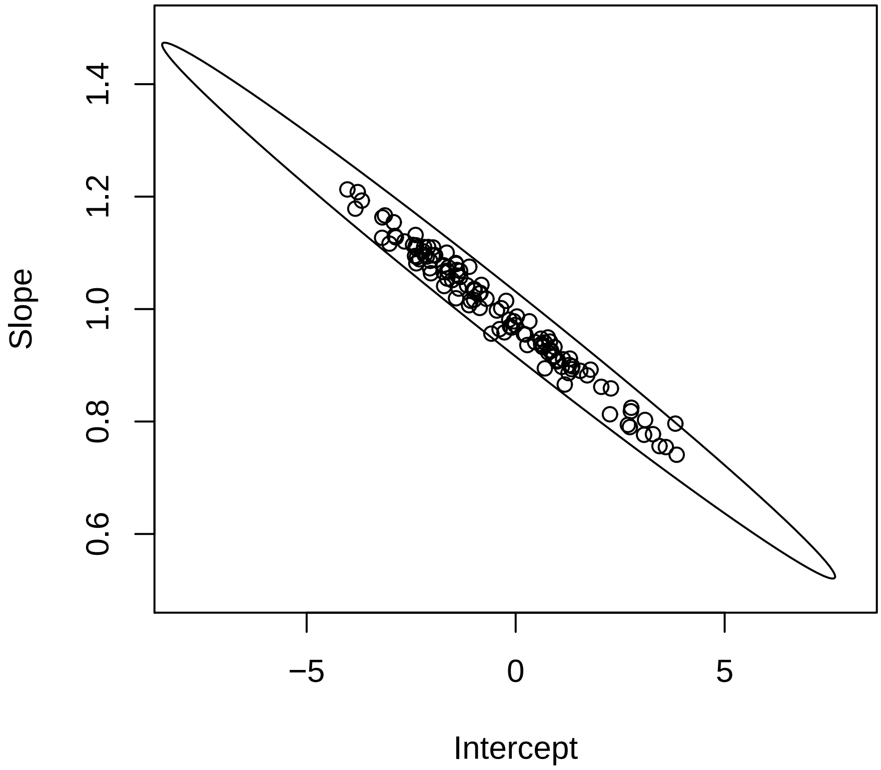

Figure 3.5 Example of a scatter plot of slope versus intercept, with a 99.9 % robust confidence ellipse.

standard deviations (SD) (s_a, s_b) and correlation r. When plotted on a two-dimensional graph, the points will be positioned in a cloud around which an ellipse-shaped confidence region can be drawn, as in Figure 3.2.

The center of the ellipse is the intercept and slope of the mean regression line and the shape is characterized by their SDs (s_a, s_b) and correlation r. A point outside the ellipse corresponds to a laboratory with significant outlying bias. The minimum covariance determinant (MCD) estimator [142, 189] is proposed as a robust measure of means, SDs and correlation of the bivariate distribution. Mahalanobis distances based on the MCD estimator are used as a measure of individual outliers and are calculated for all regression lines, also the lines excluded because of exceeding variability. Regression lines for which the Mahalanobis distance is larger than the upper α_4-percentile of a χ^2-distribution with two degrees of freedom correspond to laboratories exhibiting unacceptable bias.

3.7.4 Example

An analysis was performed on the example data set by grouping the surveys in four groups of 10 surveys (surveys 1-10, 11-20, 21-30 and 31-40). For each group of surveys, outliers were identified against the individual regression line for each laboratory as described in the first step of the algorithm. Following, the residual error was calculated by taking the square root of the mean residual variances of each regression line. The residual variances were subject to the second step of the algorithm. Finally, regression lines not exhibiting exceeding residual variance were subject to the third step. The average intercept and slope of the regression lines used in the third step was calculated. All regression lines were evaluated for bias and the percentage of bias outliers was counted. The first step was performed with α_1=0.001 and $\alpha_2 = 0.01$,the second step was performed with $\alpha_3 = 0.01$ and a trimming of 5% and the last step was performed with α_4=0.001.

Next, the analyzes was repeated without leaving out any regression outlier in the calculations of the second step and without any regression line with exceeding variability in the third step. The outliers against the regression line are found with very high probability, while the method protects well against erroneously indicating non-outliers as outliers: the negative predictive value is at least 97.3%, while the positive predictive value is at least 90.9 %. The residual error reflects very well the standard deviation of the original data: rounded to an integer, they are all equal to the original standard deviation of the data and no interference of outliers is detected.

The algorithm indicating exceeding variabilities as outliers however, lacks

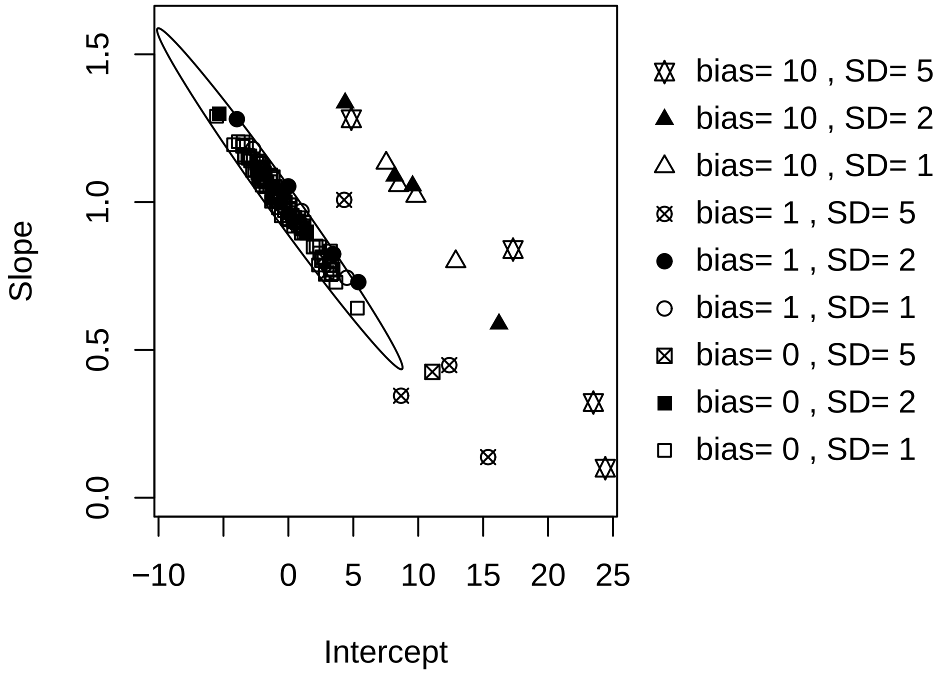

Figure 3.6 Scatter plot of slope versus intercept with a robust confidence ellipse for a selection of laboratories and surveys, for the hypothetical example, with induced bias and increased variability. Results based on 152 simulated laboratories.

some sensitivity: laboratories with a standard deviation of 2 are almost not detected as having exceeding variability; laboratories with a standard deviation of 5 on the contrary are well indicated as having exceeding variability. The estimation of the regression coefficients follows the bias of the data very well: the slope should be one and the average intercept should approach the induced bias. Only for the cases with high induced variability, the estimation is weaker. All laboratories having a bias of 10 were flagged for exceeding bias, but they were not the only ones: as soon as the variability of the data increases, the chance of being flagged for outlying bias increases as well. A scatter plot of intercept and slope values of the laboratories for which no regression outliers were induced is given in Figure 3.6. As can be seen here, all regression lines with high induced bias are outside of the confidence ellipse, as is the case for only a part of the regression lines with low induced bias.

It is interesting to know what the performance indicators would be if no val-

Table 3.8 Performance statistics of the example data set for outlier rates in the first step, residual error in the second step and bias in the third step. NPV stands for Negative Predictive Value, i.e. the percentage of correctly identified non-outlying points and PPV for Positive Predictive Value, i.e. correctly identified outlying points. Results based on 224 simulated laboratories, having reported values for 10 parameters during 40 surveys, with induced bias, increased variability and outliers.

			Regression		Mean	Variability	Average	Average	Bias
Bias	SD	Outliers	Outliers		residual	outliers	intercept	slope	outliers
		(%)	NPV (%)	PPV (%)	error	(%)			(%)
0	1	0	99.6	(.)	0.99	0	-0.50	1.0	0
0	1	5	97.3	100	0.82	0	0.49	0.96	0
0	1	10	99.3	100	0.91	0	-0.22	0.99	0
0	2	0	100	(.)	1.9	0	2.7	0.91	0
0	2	5	99.3	100	1.9	12.5	0.17	0.98	0
0	2	10	98.1	100	1.8	0	-0.1	1.0	6.30
0	5	0	99.4	(.)	4.6	93.8	6.8	0.73	43.8
0	5	5	100	100	5.2	93.8	4.3	0.89	68.8
0	5	10	100	100	4.8	93.8	2.6	0.90	43.8
1	1	0	99.4	(.)	0.98	0	0.78	1.0	0
1	1	5	99.3	100	1.1	0	2.5	0.92	0
1	1	10	99.3	100	0.95	0	-1.0	1.0	0
1	2	0	98.1	(.)	1.9	0	-0.79	1.0	6.30
1	2	5	100	100	2.2	12.5	2.0	0.97	18.5
1	2	10	100	100	2.0	6.25	0.14	1.0	12.5
1	5	0	98.1	(.)	4.9	81.25	-2.4	1.0	62.5
1	5	5	100	100	4.7	93.8	2.7	0.92	56.3
1	5	10	100	100	5.2	87.5	-7.4	1.2	62.5
10	1	0	99.4	(.)	0.97	0	10	0.98	100
10	1	5	99.3	100	0.98	0	11	0.97	100
10	1	10	100	100	1.0	0	12	0.96	100
10	2	0	99.4	(.)	2.0	6.25	13	0.92	100
10	2	5	99.3	100	2.0	6.25	8.8	1.0	100
10	2	10	100	100	1.8	0	11	0.97	100
10	5	0	99.4	(.)	4.9	93.8	6.1	1.1	100
10	5	5	100	100	4.7	93.8	11	1.0	100
10	5	10	100	90.9	4.9	87.5	15.	0.80	100

ues would have been left out in the first or second step. The results are listed in Table 3.9. Not excluding outliers against the regression line has an adverse effect on the estimators of the regression line and flagging rates for outliers or bias. Even with a modest outlier frequency rate of 5 %, almost half of the laboratories are flagged for exceeding variability and more than one third for exceeding bias. In addition, the residual error is overestimated as soon as outliers appear. The estimated mean residual error in presence of outliers is higher than the estimated mean residual error of the data simulated with the highest standard deviation and without outliers. Even more, because of the presence of regression lines of the data simulated with the highest standard deviation and without outliers is flagged for exceeding variability. In addition, outliers cause an increase of proportion of falsely flagged regression lines for bias.

3.7.5 Comments

The technique displayed here uses limits based on statistical analysis of the performance attained and is different from evaluation procedures that use limits based on experience or biological variability, although the possibility exists to introduce the latter in the second and third step of the procedure. By combining data from different samples, the method helps laboratories to understand their results in a better way and unveils more information than a parameter performance characteristic based on such as total error, such as the performance index scoring system or the mean rankings. The method's capability for searching outliers of different types is, as far as we know, a new and definitely useful method for addressing the different types of errors seen in EQA programmes. The first step excellently identifies outliers against the laboratory's individual regression line with high sensitivity and specificity, yielding two major benefits. First of all, outliers cannot be used as reliable indicators for long-term performance of the analytical phase of the Total Testing Process, but may be informative for the laboratories and support their strategies to improve the pre- or post analytical phases. Secondly, excluding outliers before estimating the regression line gives a pure estimate of the precision and trueness of each individual laboratory's analytical pro-

Table 3.9 Mean variability, intercept, slop of regression lines and mean flagging rate for variability and bias outliers if all data would have been taken into account for the hypothetical example, with induced bias, increased variability and outliers. Results based on 224 simulated laboratories.

Bias	SD	Outliers (%)	Mean residual standard error	Variability outliers (%)	Average intercept	Average slope	Bias outliers (%)
0	1	0	1.00	0	-0.5	1.00	0
0	1	5	7.91	43.8	8.0	0.88	37.5
0	1	10	9.25	62.5	-15.0	1.50	50.0
0	2	0	1.86	0	2.7	0.90	6.30
0	2	5	6.72	37.5	-0.5	0.95	43.8
0	2	10	8.03	50	3.9	0.99	56.3
0	5	0	4.74	0	6.3	0.74	43.8
0	5	5	7.93	31.3	4.7	0.90	62.5
0	5	10	11.0	81.3	3.6	0.95	93.8
1	1	0	0.99	0	0.6	1.00	0
1	1	5	7.65	56.3	-2.9	1.20	50.0
1	1	10	8.39	68.8	10.0	0.72	68.8
1	2	0	2.12	0	-2.3	1.10	6.3
1	2	5	5.49	25	0.2	1.10	37.5
1	2	10	9.88	68.8	9.0	0.96	68.8
1	5	0	5.16	0	-0.3	0.93	56.3
1	5	5	8.43	43.8	1.0	0.72	68.8
1	5	10	10.04	81.3	-10.0	1.40	81.3
10	1	0	0.99	0	10.0	0.99	100
10	1	5	5.38	12.5	2.8	1.30	100
10	1	10	6.74	31.3	1.4	1.30	100
10	2	0	2.09	0	12.7	0.93	100
10	2	5	5.08	6.3	23.0	0.72	100
10	2	10	5.51	25	12.0	1.02	100
10	5	0	5.01	0	5.5	1.07	100
10	5	5	6.35	31.3	20.0	0.81	100
10	5	10	6.57	18.8	19.0	0.71	100

cess. Outlying regression lines in the second and third steps encompass a global evaluation of the analytical phase and values exceeding the limits of acceptability that are found in these steps indicate poor laboratory analytical performance. Regression lines flagged for high variability point towards possibilities for increasing their reprodbility. Regression lines flagged for high bias point towards measurements that are constantly too high or too low for a certain range of concentrations. Care should be taken however when a laboratory finds its regression flagged for exceeding bias and variability. The high variability may yield an uncertain estimation of the regression line, that bias has become difficult to interpret. In this case, laboratories would better put efforts first in increasing their reproducibility.

3.8 Conclusion

A combined analysis of different samples or parameters informs the laboratories better about their flaws rather than a report of just one sample and it allows monitoring laboratory performance over time as well. Various methods performing this kind of analysis are available. Some of them, such as the MRVIS, mean ranking scoring, Z/U scores, all have in common to count the number of times a laboratory reports a result beyond some limits. For the MRVIS, these limits are calculated using previous results, for performance scoring and z-scores, they are defined in comparison with the results reported by the other laboratories and for the u-scores count, they are defined with respect to analytical performance goals. Other approaches, such as the LCV_a-ACD and the newly introduced 3-step method, build a regression model and derive estimates of bias and within-laboratory variability. If assigned values can be obtained with high precision, several of the methods, such as the MRVIS, Pu, LCV_a-ACD and the 3-step method can be applied for surveys with a small number of participants.

The comparison of the different approaches using the same data set is to our knowledge new. We have tried to compare them in a data set in which some specific errors were induced. Using data from real EQA rounds has the advantage that their relevance is shown for the data they have been designed, but the disadvantage is that no deliberate deviations were introduced in the data. Artificial data sets, on the other hand, give the advantage to induce specific errors and their relevance depends on how well they approach real EQA data.

The use of the artificial data set has shown that several methods do not always clearly measure what they indicate. Various techniques, such as the MRVIS or mean ranking scores, for example, were not completely able to distinguish between bias and variability. In addition, a particular flaw of the LCVa-ACD technique was observed for laboratories with a slightly positive bias: the LCVa decreases, indicating better performance, while, in fact, the analytical quality of the result decreases because of a bias. Furthermore, EQA organizers using methods that combine results from a series of different parameters should be warned that a good performance for the overall test may hide bad performance for a particular test. As a consequence, these types of tests should be utilized together with a detailed analysis for each parameter separately. Moreover, some of the methods lack any robustness. A laboratory performing well in comparison with the other laboratories and with respect to analytical goals may be flagged as soon as it reports one strongly deviating value, that does reflect any deviation in the analytical process. In addition, it should be born in mind that the nature of mistakes in the laboratory are different from phase to phase. As seen in Chapter 1, mistakes in the pre- and post analytical phase produce usually extremely deviating results, that are best described with frequency of occurrence than the size of the deviation. Errors in the analytical phase are better explained by high bias or variability.

While several methods deal with strongly deviating results by filtering them out using robust or non-parametric techniques, so far, the 3-step method is the only method that pays attention to these errors and interprets them as mistakes due to the pre- or post analytical phase. The method is also superior to the other methods for distinguishing between bias and variability. However, the method is not always applicable and is built on some assumptions that should be met before it is interpreted. It relies heavily on the correct calculation of target values. When the latter are wrongly set, the frequency of outliers in the first step may increase and the variability estimate may be underestimated. In addition, it assumes a linear relation between target and reported values. Whenever a laboratory finds that its outliers against te regression model are concentrated around a particular concentration range, mostly near lower or higher concentrations, it should focus rather on the linearity of its results - so more on the analytical phase - than searching for reducing errors in the pre- or post-analytical phase. Also, EQA organizers that find a high occurence of regression outliers for a particular sample should consider that the sample may suffer from unforeseen matrix effects and redraw it for analysis.

CHAPTER 4

Applications of the 3-step method

4.1 Introduction

The 3-step procedure described in Chapter 3 can be used under various conditions. It can be applied to any EQA data set obtained for a series of samples where a target value for one or more parameters can be set. A major prerequisite is the absence of matrix effects and the assumption of a linear relation between target values and reported values. The approach may reveal information about the frequency of accidental mistakes, within-laboratory variability and bias.

When data are obtained over a longer period of time, the variability and bias can be considered as a long-term analytical variability and bias. This information is useful in addition to the information obtained during the validation protocol of an analytical process, since it is usually recorded over a much longer time range than is the phase for a validation process. Moreover, the time range can be subdivided into different distinct periods and a comparison between periods may reveal information about the evolution of quality in the clinical laboratory.

When data are obtained from different samples sent in the same survey, the method reveals information about the within-laboratory standard deviation and bias for the particular day of analysis and the results should be comparable with those of the method validation. For the EQA organizer, the method may add useful information when the performance of laboratories and/or methodologies is assessed over time and or between methods. Individual laboratory performance may be monitored and a method assessment can be made over time by comparing the performance over different sub-periods. In this chapter, several data sets are analyzed with the 3-step procedure. The method was applied to data obtained in four different surveys that were organized by the Belgian EQA scheme.

4.2 Applying the 3-step method

The analysis proceeds each time according to the structure described below. For each laboeratory a regression line is calculated and outliers against it are identified. Then the regression residual variabilities of all regression lines are compared with each other and afterwards the bias, reflected by the position of the regression lines, is evaluated.

4.2.1 Identification of outliers

Data were grouped per laboratory. When different methodologies were used within the laboratories or when different parameters were analyzed, the grouping was made per laboratory, applied methodology and parameter. When different periods were involved in the analysis, one regression line was constructed spanning over all time periods.

The outliers against each regression line were identified with α_1 set at 0.001 and α_2 at 0.01 and their frequency was analyzed per laboratory, per methodology, per parameter, per period, or any combination of these. Frequencies between periods or methodologies were compared with a Chi-square statistic or, in case few data are available, with the Fisher's exact test and/or visualized by a histogram. A generalized linear model (GLM) approach to assess differences in outlier frequencies between methodologies or periods was applied as well.

Since regression outlier calculation encompasses a comparison within each individual laboratory's results, all results were interpreted, also the results obtained by methods used by one or a few laboratories. The only inclusion criterion that was applied was a minimal number of data for each regression line.

4.2.2 Identification of exceeding variability

For each regression line, the method allows to assess whether the regression line's variability exceeds a certain limit; the parameter α_3 was set at 0.01 and a trimming of 10% was applied. If applicable, the frequency of regression lines with exceeding variability was compared between methodologies and/or periods with the Fisher's exact test. In addition, the residual standard error of the regression lines, calculated without taking into account the exceeding residual standard error, was considered as a representative statistic for the state of the art analytical variability of a certain parameter, method and/or period.

The results of methods that were applied by at least 4 (or, in one study, 6) laboratories throughout the study period were compared between time periods and methods with a general linear mixed model (GLMM), in which the methods and periods were considered as fixed factors and laboratories as a random factor. When no interaction between periods and methods were significant, a comparison was made between periods after averaging results over all methods and between methods after averaging results over all periods. In case of a significant interaction, periods were compared per method and methods per period. Comparisons were each time adjusted for simultaneous hypothesis testing according to Tukey or Sidak.

4.2.3 Identification of exceeding bias

Similarly to the evaluation of regression lines with exceeding variability, regression lines with exceeding bias may be identified as well and the frequency of regression lines with exceeding bias can be compared between methods and/or periods via the Fisher's exact test. To identify regression lines with exceeding bias, the parameter α_4 was set at 0.001.

In addition, the difference between a regression line and the 45°-line may be calculated by taking the maximum distance between the 45°-line and the regression line within the measurement range of the data under consideration. This distance may be expressed as a maximum bias, (absolute value of maximal distance between the regression line and the 45°-line), or as relative bias (maximal distance divided by corresponding target value). Both absolute and relative biases are continuous and, as well as residual standard errors, may be compared between periods and/or methods via a GLMM and subsequent multiple comparisons between groups. The results of some methods may show a high absolute bias and a low relative bias, or vice versa. The first case means that the regression lines deviate from the 45°-line for higher target values, the later means that the regression lines deviate from the 45°-line for the lower target values. Analogous to the comparison of residual standard error, regression lines with exceeding variability and bias should be excluded for calculation.

4.3 Belgian EQA for ethanol determination

4.3.1 Introduction

Assay of ethanol in blood is a widely applied technique for toxicological and forensic purposes. The first application is often used in the emergency laboratory for diagnosing acute intoxication; it is performed under strict conditions and serves among others to deliver evidence for legal cases. Thus, laboratories often have two methods at their disposal: one method for routine intoxication measurement and one routine for medico-legal purposes. The latter, in particular, needs to be an accurate, precise and selective procedure for determining blood alcohol. Chromatographic methods are best suited here and, in Belgium, required for medico-legal reasons. They enable a reliable, fast, precise and sensitive analysis. Other methods are often based on the alcohol-dehydrogenase-facilitated production of NADH+ with ethanol as substrate and subsequent colorization of a dye via diaphorase. Subsequently, the colored dye is spectrophotometrically measured. This reaction is not entirely selective for blood ethanol. There is a high interference with other alcohols, such as methanol or isopropanol. Also, endogenous dehydrogenases and substrates other than alcohols, such as lactate, may interfere with NADH production, leading to falsely elevated ethanol measurements. Often, laboratories use both a chromatographic and ADH-based method.

The Belgian EQA for blood or serum ethanol has been running since 2002. The sample material is generally prepared from fresh serum samples, spiked with ethanol and stored deep-frozen till send-out. Participants are requested to keep the vials cooled till analysis. Data were reported in an electronic form starting from 2003 and laboratories were able to register measurements with an enzymatic-based and a chromatographic method together.

A selection was made to obtain samples solely spiked with ethanol with a minimal concentration of 0.1 g/L for a long-term follow up. Only exact results were taken into account, i.e. any censored result reported like '<x' or '>x' was discarded. To evaluate performance over time, the time range was arbitrarily divided into three distinct periods: 2003-2005, 2006-2008 and 2009-2011.

4.3.2 Results

The study entails 15791 reported results of 79 samples by 218 laboratories over a time span of 9 years. The number of data, samples and laboratories are listed in Table 4.1. A total of 158 laboratories has reported results over the whole study period, of which 123 (78 %) laboratories have been using the same methodology for the whole period.

Table 4.1 Number of data for long-term alcohol study of the results of the Belgian EQA for blood alcohol testing.

Period	No. of observations	No. of samples	No. of laboratories
2003-2005	3241	28	193
2006-2008	3206	26	195
2009-2011	3174	25	184

Outliers against the regression line

Outliers against the laboratories' individual regression lines were first counted per period and laboratory and their frequency was displayed by histograms (Fig 4.1). There was a tendency towards less outliers per laboratory for the period 2008-2011. A generalized linear mixed model showed that there was a significant drop in the outlier frequency between this and the two former periods.

The outlier frequency counted per method and period is shown in Table 4.2. The total number of outliers decreased significantly from the first two towards the last period. When counted per method, this drop was significant for the ADH-methods of Vitros, Roche and Dade (Emit) and for the headspace chromatographic method.

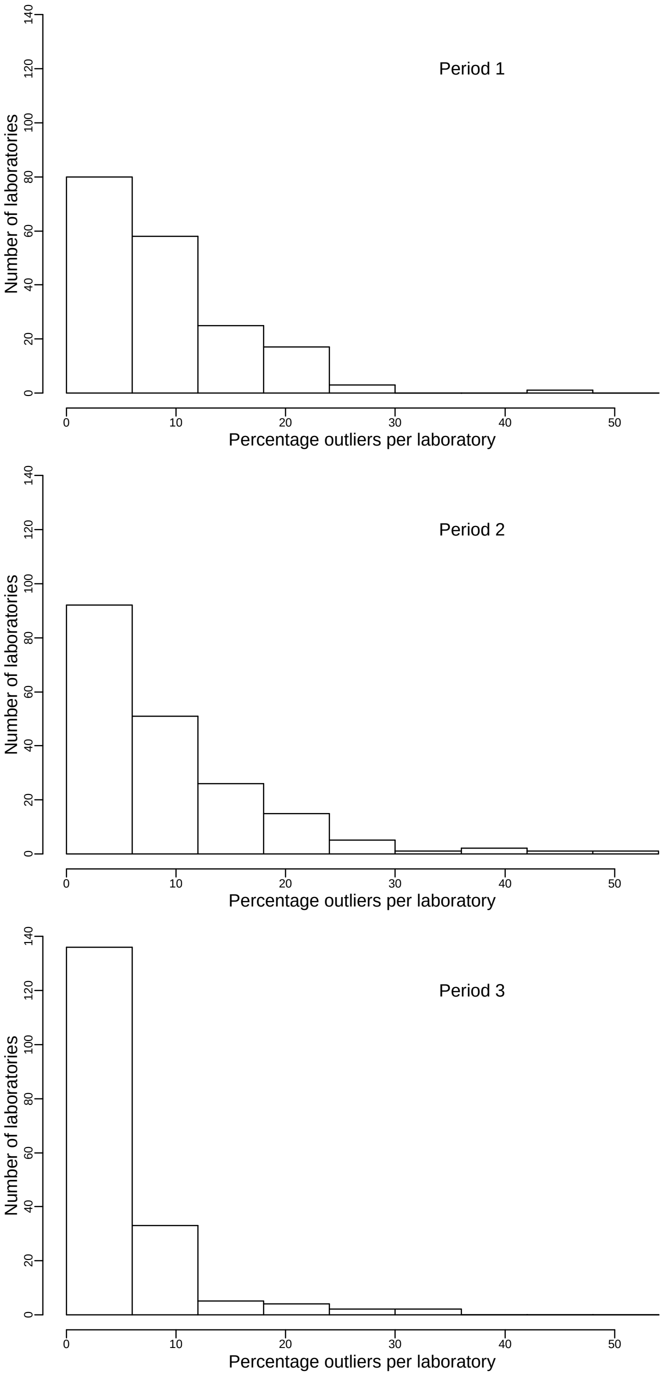

Figure 4.1 Histograms of percentage of regression outliers per laboratory, for the three time periods in the long-term alcohol study of the results of the Belgian EQA for blood alcohol testing. Period 1: 2003-2005, Period 2: 2006-2008, Period 3: 2009-2011.

Table 4.2 Outliers against the individual regression line for a certain method applied by the laboratories individually for the long-term alcohol study of the results of the Belgian EQA for blood alcohol testing. The percentages of outliers are shown followed by the total number of results for the corresponding method and period. N stands for the total number of results for a certain period and method. Period 1: 2003-2005, Period 2: 2006-2008, Period 3: 2009-2011.

	Period 1		Period 2		Period 3		
Method	N	Outliers (%)	N	Outliers (%)	N	Outliers (%)	
ADH- Abbott AxSym	46	0	46		50	4.00	
ADH- Abbott TDx/ADx	33	3.03	30	20.0	14	0	
ADH- Beckman	38	5.26	52	17.3	52	5.77	
ADH - Dade Dimension	56	14.3	52	3.85	46	6.52	
ADH- Dade (Emit)	168	9.52	148	8.11	155	0.65	*
ADH- Roche	1903	8.57	1860	6.29	1880	3.35	*
ADH- Vitros	580	12.8	564	17.7	541	2.96	*
Direct Gas Chromatography (capillary-column)	84	23.8	96	3.12	104	3.85	*
Direct Gas Chromatography (packed-column)	149	5.37	155	9.03	150	3.32	
Headspace Chromatography (capillary-column)	184	11.4	203	8.37	182	3.85	*
Total	3241	9.66	3206	8.73	3174	3.28	*

*A significantly lower frequency was observed in period 3

Variability around the linear regression lines

The frequency of regression lines with exceeding bias, counted per method and period and the regression residual standard error are shown in Table 4.3. The total number of regression outliers per period didn't differ significantly from each other, neither was there any method for which a significant evolution in residual variability outlier rate was detected. Averaged over the whole study period, 6.4 % of the laboratories were flagged for exceeding variability.

For a comparison between methods, the methods used by at least 4 for all the periods laboratories were considered. For none of the methods, a significant evolution over time was recorded, nor an interaction between the time and method factor. There was however a significant difference between the mean residual standard error values for the methods: the Direct Gas Chromatography had a significantly lower variability than the other methods, except for the Headspace chromatography. The latter itself had an intermediate variability, which was not different from the Direct Gas chromatography nor from the enzymatic methods.

Table 4.3 Frequency of variability outliers (in percentage) and regression residual standard error for the most popular methods over the whole time period in the long-term alcohol study of the results of the Belgian EQA for blood alcohol testing. Period 1: 2003-2005, Period 2: 2006-2008, Period 3: 2008-2011.

Method	N	Regression lines with exceeding variability (%)			Regression residual standard error			
		Period			Period			
		1	2	3	1	2	3	
ADH- Abbott AxSym	2	100	0	50				
ADH- Abbott TDx/ADx	1	100	100	100				
ADH- Beckman	2	50	0	0				
ADH - Dade Dimension	2	0	0	50				
ADH- Dade (Emit)	6	0	0	16.7	0.024	0.028	0.03	
ADH- Roche	72	2.78	8.33	1.39	0.025	0.022	0.022	
ADH- Vitros	22	4.55	18.2	4.55	0.024	0.028	0.018	
Direct Gas Chromatography (capillary-column)	3	0	0	0				
Direct Gas Chromatography (packed-column)	6	0	0	0	0.014	0.016	0.017	*
Headspace Chromatography (capillary-column)	7	0	0	0	0.022	0.023	0.02	
Total	123	5.69	8.94	4.88				

*Method with significantly lower regression residual standard error, except when compared to Headspace Chromatography (capillary-column)

Bias of the regression line

The frequency of regression lines that were flagged for exceeding bias, calculated using the laboratories that have used the same methodology throughout the study period, together with the maximal absolute and relative bias, are given in Table 4.4. The frequency of bias outliers is remarkably lower than the frequency of variability outliers, the totals for each period are not significantly different. Further, there was no significant evolution observable over time and the low frequency did not allow a reliable comparison between methods. Averaged over all periods, 1.1 % of the laboratories are flagged for exceeding bias.

There was neither a significant interaction nor a significant difference in absolute or relative bias between the time periods. The absolute bias, however, was higher for the ADH-method with Vitros than for other methods. The bias of this method is visualized in Figure 4.2. All the regression lines are close to the 45°-line for small values. For larger values, however, the lines tend towards values under the 45°-line. This means that, on average, results obtained with this method are slightly underestimated for larger concentrations. The bias is however so small that not any laboratory using this methodology was flagged for exceeding bias.

Table 4.4 Percentage of laboratories flagged for exceeding bias and absolute and relative bias for the most popular methods in the long-term alcohol study of the results of the Belgian EQA for blood alcohol testing. Period 1: 2003-2005, Period 2: 2006-2008, Period 3: 2008-2011.

Method	N	Bias outliers (%)				Absolute bias			Relative bias		
		Period				Period			Period		
		1	2	3		1	2	3	1	2	3
ADH- Abbott AxSym	2	0	0	0							
ADH- Abbott TDx/ADx	1	0	0	100							
ADH- Beckman	2	0	0	0							
ADH - Dade Dimension	2	0	0	0							
ADH- Dade (Emit)	6	0	16.7	0		0.05	0.14	0.03	9.2	8.1	8.0
ADH- Roche	72	1.39	0	0		0.05	0.10	0.06	13	18	12
ADH- Vitros	22	0	0	0	*	0.17	0.19	0.09	9.8	8.8	12
Direct Gas Chromatography (capillary-column)	3	33.3	0	0							
Direct Gas Chromatography (packed-column)	6	0	0	0		0.04	0.05	0.02	8.8	4.4	8.6
Headspace Chromatography (capillary-column)	7	0	0	0		0.07	0.07	0.04	12	11	9.6
Total	123	1.62	0.813	0.813							

*ADH-Vitros had a significantly higher absolute bias than the other methods

4.3.3 Discussion

The average variability and bias of the reported results has remained stable over time. A slight difference between methods is visible, with mainly the Vitros method showing a slight negative bias with increasing alcohol concentration. In addition, the frequency of laboratories found to underperform with respect to bias or variability remained stable over time and the main problem is caused by exceeding variability.

In contrast to a stable variability and bias, a clear drop in accidental mistakes has been recorded since 2009. An explanation may lay in the fact that from 2010 on, a post-analytical control step was introduced to verify and eventually correct for largely exceeding values. This result points rather to a decrease in post-analytical error, partially due to the complexity of reporting results to the EQA organizer.

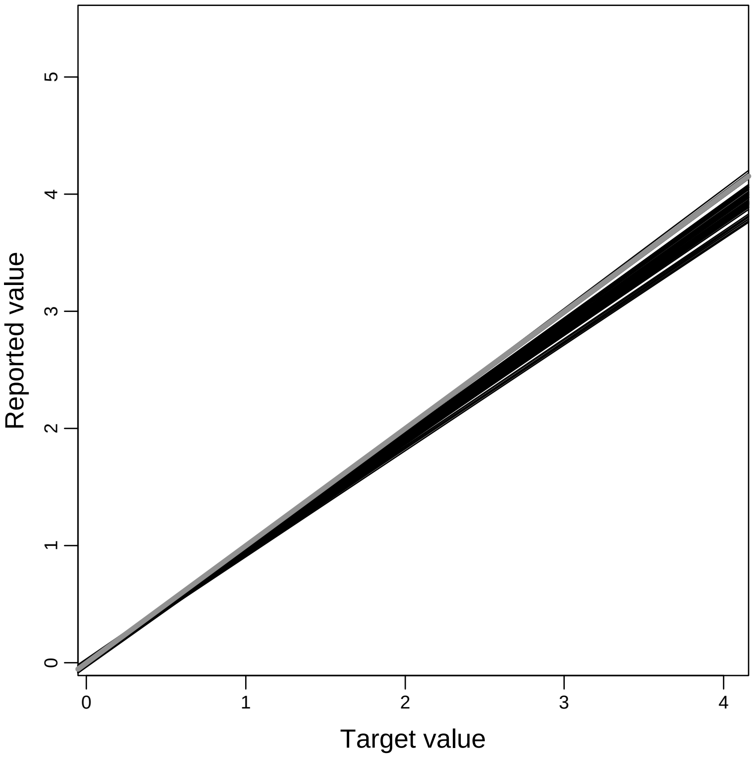

Figure 4.2 Regression lines for the results obtained for the Vitros method in the long-term alcohol study of the results of the Belgian EQA for blood alcohol testing, all periods together. The thick grey line is the 45°-line.

4.4 EQA survey for lymphocyte subset counting

4.4.1 Introduction

Flow cytometry is a widely used laboratory method for the identification and quantification of lymphocyte subsets. It is particularly helpful in the diagnosis and monitoring of cellular immunodeficiency diseases, leukaemia and lymphomas. Lymphocyte subset count by flow cytometry has been subject to external quality control procedures for more than 20 years.

The Belgian EQA scheme for lymphocyte subset counting has been running since 2000. For each send-out, fresh human peripheral blood was obtained by voluntary donation after consent from the individual donors. Aliquots of 3 ml were prepared, packaged and sent on the day of collection to the participating laboratories by overnight mail. All laboratories were encouraged to perform sample testing as soon as possible and to process the samples according to their usual procedures. Three surveys take place every year, each incorporating three samples. To avoid sample deterioration effect, only results from assays carried out within two days after collection were taken into account.

Several lymphocyte subset counts were requested, of which the percentages and absolute counts of CD3+, CD4+, CD8+, CD19+ and NK lymphocytes have been asked for almost the whole period. Since the applied methodology is rather diverse, the data were analyzed without taking into account differences between methodology.

4.4.2 Results

A summary of the number of data involved for each period is given in Table 4.5. For a follow-up between the distinct periods, only results of laboratories that have reported results over the whole period were taken into account.

Table 4.5 Number of data, samples and laboratories for the long-term lymphocyte subset counting follow-up. Period 1: mid 2000-beginning 2004, Period 2: mid 2004-end 2007, Period 3: beginning 2008-mid 2011.

Parameter	No. laboratories,	Period 1			Period 2			Period 3		
	all periods	N	Samples	Laboratories	N	Samples	Laboratories	N	Samples	Laboratories
CD3 %	43	1419	32	56	1613	33	58	1458	33	50
CD3	43	1284	32	54	1552	33	58	1381	30	50
CD4 %	43	1424	32	56	1613	33	58	1458	33	50
CD4	43	1309	32	54	1553	33	58	1282	30	50
CD8 %	43	1424	32	56	1613	33	58	1452	33	50
CD8	43	1289	32	54	1553	33	58	1381	30	50
CD19 %	43	1408	32	56	1604	33	57	1429	33	50
CD19	43	1270	32	54	1544	33	57	1355	30	50
NK %	38	512	12	45	1556	33	56	1418	33	49
NK	38	488	12	45	1502	33	56	1348	30	49

Outliers against the regression line

The percentage of outliers per laboratory, registered for all parameters together, is shown in Figure 4.3. The frequency of outliers was significantly lower in period 2 with respect to the two other periods, with a median number of regression outliers was 3 percent in the former and 5 % in the latter. All except one laboratory had at least one regression outlier for at least one parameter.

The same data, grouped for all laboratories together and split per parameter and period, are shown in Table 4.6. The outlier rate merged over all parameters dropped from the first to the second and increased again in the third period. Considered per parameter, the outlier rate was highest in the first period and dropped significantly for percentages of CD3, CD4 and CD8 and absolute counts of CD3 and CD8, while the rate from 2008 to 2011 had an intermediate value. The outlier rate rose significantly between period 2 and 3 for percentages and absolute counts of CD19.

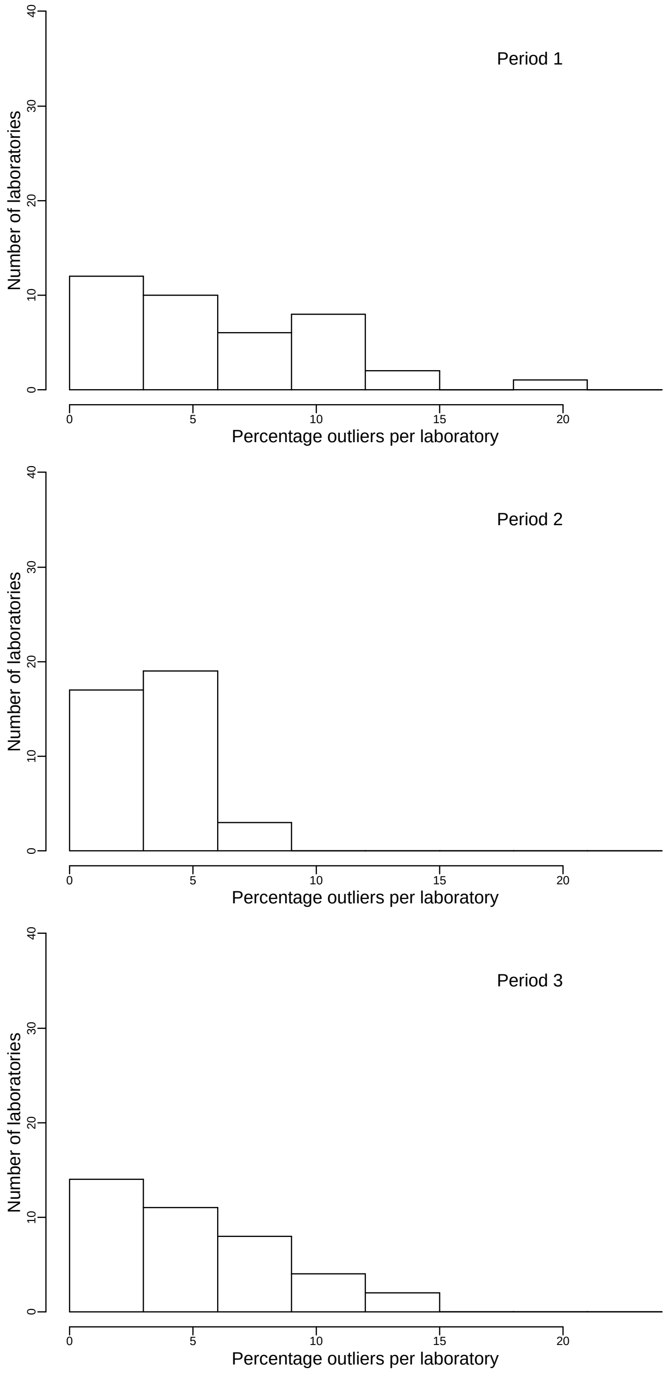

Figure 4.3 Distribution of outliers against the individual regression lines for the lymphocyte subset counting follow-up, considered over all parameters, for the three distinct periods. Period 1: mid 2000-beginning 2004, Period 2: mid 2004-end 2007, Period 3: beginning 2008-mid 2011.

Table 4.6 Regression outliers against the regression model, merged over all laboratories and displayed per year and parameter for the lymphocyte subset counting EQA follow-up, for the three distinct periods. Period 1: mid 2000-beginning 2004, Period 2: mid 2004-end 2007, Period 3: beginning 2008-mid 2011

	Period 1		Period 2		Period 3		
Parameter	N	Outliers (%)	N	Outliers(%)	N	Outliers(%)	
CD3 %	1058	6.90	1236	3.56	1159	4.75	(a)
CD3	961	6.04	1157	2.68	1061	3.58	(a)
CD4 %	1060	7.36	1236	3.56	1159	5.61	(a)
CD4	995	5.23	1191	3.02	1091	5.04	
CD8 %	1060	7.45	1236	3.32	1155	4.24	(c)
CD8	975	6.15	1191	2.44	1091	4.77	(a)
CD19 %	1056	4.92	1269	3.63	1182	6.94	(b)
CD19	960	5.00	1188	4.04	1091	7.79	(b)
NK %	228	0.877	621	4.03	578	6.23	
NK	168	2.98	450	5.11	405	6.17	
Total	8521	5.95	10775	3.41	9972	5.44	

(a) Periods 1 and 2 differed significantly; (b) periods 2 and 3 differed significantly; (c) period 2 significantly different from periods 1 and 3

Variability around the linear regression lines

The percentage of regression lines with exceeding bias and their residual standard error are shown in Table 4.7. Except for NK, absolute counts show a higher rate of exceeding values for regression variability (P<0.001). There was no evolution over time visible for any of the parameters.

Considering the number of regression lines with exceeding variability for each laboratory, more than 80 percent of the laboratories were not flagged for exceeding variability for any parameter, 34 in period 1, 32 in period 2 and 33 in period 3. The worst performers showed in period 1 50 %, in period 2 37.5 % and in period 1 40 % regression lines with exceeding variability. Considering the number of regression lines with exceeding bias for each laboratory, more than 60 percent of the laboratories were not flagged for exceeding variability for any parameter, 24 in period 1, 25 in period 2 and period 3. The worst performers showed a frequency of regression lines with exceeding variability of 87.5 % in period 1 and 50 % in periods 2 and 3.

Per parameter, the mean values for the three periods are very similar and for none of the parameters, a significant difference between the periods exists.

Table 4.7 Frequency of regression lines with exceeding variability and mean regression residual standard error for the three distinct periods, calculated per parameter for the lymphocyte subset counting EQA follow-up, for the three distinct periods. Period 1: mid 2000-beginning 2004, Period 2: mid 2004-end 2007, Period 3: beginning 2008-mid 2011.

Parameter	N	Lines with exceeding variability			Regression residual standard error		
		Period			Period		
		1	2	3	1	2	3
CD3 %	38	7.90	2.63	2.63	1.8	1.7	1.7
CD3	37	10.8	10.8	13.5	0.086	0.087	0.084
CD4 %	38	2.63	5.26	10.5	1.7	1.70	1.6
CD4	38	13.2	21.1	18.4	0.061	0.057	0.061
CD8 %	38	10.5	5.26	10.5	1.4	1.3	1.3
CD8	38	15.8	7.90	10.5	0.040	0.037	0.04
CD19 %	39	7.69	7.69	12.8	0.80	0.77	0.81
CD19	38	15.8	7.90	18.4	0.018	0.018	0.02
NK %	19	5.26	10.5	10.5	1.1	0.96	0.83
NK	14	0	7.14	7.14	0.025	0.024	0.024

Bias of the regression line

The number of lines identified with exceeding bias have been counted for each parameter and are shown per period in Table 4.8. The numbers were very low and there was no significant evolution over time.

Remark that slightly higher numbers for exceeding bias were found for the percentages than for the absolute counts. Only 2 laboratories were flagged for exceeding bias for four or more parameters. Both of them were flagged

Table 4.8 Count of lines with exceeding bias, per parameter and period for the lymphocyte subset counting EQA follow-up, for the three distinct periods. Period 1: mid 2000-beginning 2004, Period 2: mid 2004-end 2007, Period 3: beginning 2008-mid 2011.

Parameter	No of	Number of bias outliers		
	laboratories	Period 1	Period 2	Period 3
CD3 %	37	1	1	1
CD3	37	0	0	0
CD4 %	37	1	0	2
CD4	38	0	0	0
CD8 %	35	3	2	3
CD8	37	1	0	0
CD19 %	38	1	2	4
CD19	35	3	1	0
NK %	18	1	1	0
NK	14	0	1	0
Total	326	11	8	10

for exceeding variability as well. The mean absolute and relative bias were calculated for each parameter per period and are displayed in 4.9.

For the percentage counts, relative bias was lower for CD3 % and CD4 % and absolute bias was significantly higher for only CD3 %.

For the absolute counts, all absolute bias measures differed significantly from each other, except for NK, that had an intermediate value CD8 and CD19. Also the relative bias values differed significantly from each other, except for CD4, that had an intermediate value between CD3 and CD8.

No evolution in mean absolute or relative bias was observed over time.

4.4.3 Discussion

By far, the largest deviations from the ideal behavior is found by the identification of outliers against the individual regression line of each laboratory, for each parameter. Discarding outliers, variability and bias remained markedly stable over time. A large majority of laboratories was not flagged for regression variability or bias, while a few poor performers were identified with

Table 4.9 Relative and absolute bias, calculated per parameter for each period separately in the lymphocyte subset counting EQA follow-up. Period 1: mid 2000-beginning 2004, Period 2: mid 2004-end 2007, Period 3: beginning 2008-mid 2011.

Parameter	Absolute bias			Relative bias		
	Period			Period		
	1	2	3	1	2	3
CD3 %	1.5	1.60	1.46	0.021	0.021	0.019
CD4 %	1.1	1.49	1.43	0.026	0.031	0.032
CD8 %	1.6	1.52	1.52	0.064	0.079	0.054
CD19 %	1.0	1.16	1.13	0.087	0.10	0.087
NK %	1.0	1.26	0.96	0.12	0.17	0.17
CD3	0.11	0.076	0.086	0.067	0.056	0.060
CD4	0.068	0.063	0.072	0.07	0.066	0.090
CD8	0.040	0.049	0.055	0.10	0.11	0.087
CD19	0.026	0.031	0.031	0.15	0.15	0.11
NK	0.031	0.033	0.032	0.17	0.193	0.19

up to 50 % of their regression lines with exceeding bias and/or variability. These findings are in line with those published by Levering, who also found a modest improvement in laboratory performance.

Further, this study shows that the evolution of performance was not a constantly improving process. For different parameters, a worsening was observed between the second and third period. This shows that EQA organizers should remain alert for changing performance and never cease to trust that once a quality level has been reached, performance will never drop again. Even more, the method, although used here in a static way, with distinct time periods and a strict selection of data, may be applied dynamically as well, with moving time windows of variable length. When used by the individual laboratories, they may help in quickly identifying deviating behavior and characterizing the source of errors.

We have tried to identify the cause of the problems of bias of several laboratories by presenting several cases to the EQA's expert committee for lymphocyte subset counting. The bias of one laboratory for CD3+, CD4+ and CD8+ cells could probably be attributed to the extreme susceptibility of the Cell-Dyn Sapphire haematology analyser for sample ageing. For another laboratory, the negative bias found for the determination of CD19+ lymphocytes could be attributed to an inadequate setting of the light-scatter gate leading to the presence of too much nonlymphocyte contaminants causing an underestimation of the true percentage of CD19+ cells.

4.5 EQA survey for Lithium

4.5.1 Introduction

Lithium (Li^+) is a monovalent cation with antipsychotic activity and is used in the prophylaxis and the treatment of manic-depressive illnesses. Blood levels need to be closely monitored, since it has a narrow therapeutic index.

The Belgian EQA for Therapeutic Drug Monitoring assessed the performance of lithium assays in Belgium and organized in 2010 a survey involving nine fresh serum samples, with lithium concentrations ranging from 0 to 3.975 mmol/L. Some laboratories used the reference method (Flame photometry - with internal standard) and the median values of their reported values were taken as target value.

The data were obtained in one single survey and as a consequence, the performance assessment of laboratories and methods is a snapshot of the performance on the day of analysis and not a long-term evaluation. The residual standard error is more linked to a within-day repeatability in this case. One sample was discarded for the application of the 3-step procedure since it didn't contain any lithium.

The regression outlier detection step was slightly modified for analyzing this data set. A residual dot plot of several regression lines obtained in this study indicated that the regression models were heteroscedastic, with a higher variability for samples with a higher concentration of Li^+. For this reason, the following modifications were applied:

A variability measure for each sample was calculated by selecting the methods used by 6 laboratories or more. For each sample, the variance was calculated for each method after excluding outliers according to the sequential Grubbs method with an alpha of 0.05, as described in Chapter 1. The variance values obtained for the different methods were pooled per sample using the following formula:

$$\mathrm{s}^2_{\mathrm{pooled}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{k}}(\mathrm{n_i} - 1)\mathrm{s_i^2}}{\mathrm{N} - \mathrm{k}}$$

with k the number of methods, N the total number of data obtained for all the methods together and $\mathrm{n_i}$ and $\mathrm{s_i^2}$ the number of data and variance for a particular method. Next, the robust regression model was applied as usual and spurious points were identified. The last part of the regression outlier finding step, where every individual point is evaluated with respect to a simple linear regression model built without the suspect points, was built using the inverse variances for each sample as weights. Further, due to the low number of data available for each regression line and hence a low certainty about the variability and bias estimates of the regression lies, we opted in this case to compare regression lines used by at least 6 laboratories.

4.5.2 Results

The 8 samples were analyzed by 114 laboratories. The distribution of the analytical methods is shown in Table 4.10.

Outliers against the regression line

In total, the vast majority of the laboratories (108; 95%) did not produce any spurious result at all. Four laboratories had one regression outlier (one used Flame photometry with internal standard and one used Spectrophotometry-Abbott Architect/Aeroset) and 2 laboratories had 2 regression outliers (one used Atomic absorption photometry and one used Roche-Integra)

Table 4.10 Number of laboratories for each applied methodology in the 2010 Belgian EQA survey for Lithium.

Method	No. of laboratories
Atomic absorption photometry	6
Colorimetric reflectometry - OCD	22
Direct potentiometry - ISE - AVL	1
Direct potentiometry - ISE - Instr. Laboratory	2
Direct potentiometry - ISE - Menarini Spotlyte	2
Direct potentiometry - ISE - Roche - Integra	13
Enzymatic method- Spectrophotometry - Diazyme	1
Flame photometry - with internal standard	8
Flame photometry - without internal standard	2
Inductively Coupled Plasma-Mass Spectrometry	1
Spectrophotometry - Abbott - Architect / Aeroset	4
Spectrophotometry - Roche - cobas c 501	31
Spectrophotometry - Siemens - Advia	3
Spectrophotometry - Siemens Dimension	7
Spectrophotometry - Thermo electron corporation	11
Total	114

Variability around the regression line

The distribution of the regression lines with exceeding bias for the different methods and the residual standard error for the methods used by at least 4 laboratories is shown in Table 4.11. For 3 methods (Direct potentiometry-ISE, AVL and Instrument Laboratory and Enzymatic method - Spectrophotometry - Diazyme), all participating laboratories were flagged for exceeding variability.

Other laboratories flagged in this step applied Atomic absorption photometry, colorimetric reflectometry - OCD or Spectrophotometry, Siemens Dimension or Thermo electron corporation. The low numbers however didn't allow to give any statistical significance to the data. The P-value for the Fisher's exact test for difference in outlier rate per method was 0.073.

Concerning residual standard error, a significant difference between methods was found. The two methods with the lowest variability were Direct potentiometry from Roche/Integra and the Flame photometry with internal standard. Their residual standard error was significantly lower than the method with highest mean residual standard error, the colorimetric reflectometry from OCD. Other methods didn't differ significantly.

Table 4.11 Overview of regression lines with exceeding variability, for methods used by at least 6 laboratories in the 2010 Belgian EQA survey for Lithium.

Method	N	Outliers (%)	Residual standard error
Atomic absorption photometry	6	16.7	0.01
Colorimetric reflectometry - OCD	22	9.09	0.015
Direct potentiometry - ISE - AVL	1	100	
Direct potentiometry - ISE - Instr. Laboratory	2	100	
Direct potentiometry - ISE - Menarini Spotlyte	2	0	
Direct potentiometry - ISE - Roche - Integra	13	0	0.01
Enzymatic method- Spectrophotometry - Diazyme	1	100	
Flame photometry - with internal standard	8	0	0.009
Flame photometry - without internal standard	2	0	
Inductively Coupled Plasma-Mass Spectrometry	1	0	
Spectrophotometry - Abbott - Architect - Aeroset	4	0	
Spectrophotometry - Roche - Cobas c 501	31	0	0.012
Spectrophotometry - Siemens- Advia	3	0	
Spectrophotometry - Siemens - Dimension	7	14.3	0.013
Spectrophotometry - Thermo electron corporation	11	18.2	0.014
Total	114	8.77	

Note: the residual standard error of Colorimetric reflectometry - OCD was significantly higher in comparison with Direct potentiometry - ISE - Roche and Integra and Flame photometry - with internal standard

Bias of the regression line

The distribution of the frequency of regression lines with exceeding bias, together with the absolute and relative bias for the most popular methods, is shown in Table 4.11. Five laboratories were flagged for exceeding variability; they used atomic absorption photometry, direct potentiometry from AVL, spectrophotometry from Diazyme or Roche/Cobas c501, flame photometry with internal standard. A Fisher's exact test between the most popular methods didn't reveal any significant difference in outlier rate between the methods.

The method with the highest absolute bias was the spectrophotometric method from Thermo electron. It had an absolute bias that was significantly higher than all the other methods, except the Atomic absorption photometry. Relative bias was not significantly different between the methods.
Figure 4.4 explains why the Thermo electron method had a significantly higher absolute bias, but no significantly higher relative bias. All regression lines followed the 45°-line very well for lower concentrations, meaning that their intercept was close to zero. For increasing concentrations, however, the regression lines deviated from the 45°-line, meaning that they all had slopes lower than 1.

4.5.3 Discussion

A large series of different methods are used in Belgium for assaying lithium. The low frequency of outliers against the regression line indicate that laboratories deliver results which are in accordance with their own results. When the results of different laboratories are compared with each other, we can see that a relatively large amount of laboratories (8.8 %) have an analytical variability that can be considered as exceedingly large and a smaller amount of laboratories (4.4 %) show an exceeding bias. For the methods that were reported by a sufficiently large number of laboratories, a comparison of regression residual standard error and bias indicated that some methods have a significantly higher variability than others and one method has a signifi-

Table 4.12 Frequency of regression lines with exceeding bias in the 2010 Belgian EQA survey for Lithium, counted per method.

Method	N	Outliers (%)	Absolute bias	Relative bias
Atomic absorption photometry	6	16.7	0.2143	0.1156
Colorimetric reflectometry - OCD	22	0	0.2060	0.1404
Direct potentiometry - ISE - AVL	1	100		
Direct potentiometry - ISE - Instr. Laboratory	2	0		
Direct potentiometry - ISE - Menarini Spotlyte	2	0		
Direct potentiometry - ISE - Roche - Integra	13	0	0.1703	0.0839
Enzymatic method- Spectrophotometry - Diazyme	1	100		
Flame photometry - with internal standard	8	12.5	0.0754	0.0777
Flame photometry - without internal standard	2	0		
Inductively Coupled Plasma-Mass Spectrometry	1	0		
Spectrophotometry - Abbott - Architect / Aeroset	4	0		
Spectrophotometry - Roche - cobas c 501	31	3.23	0.2065	0.1226
Spectrophotometry - Siemens- Advia	3	0		
Spectrophotometry - Siemens - Dimension	7	0	0.1440	0.1723
Spectrophotometry - Thermo electron corporation	11	0	0.5399 *	0.1576
Total	114	4.39		

*Absolute bias of Spectrophotometry - Thermo electron corporation was significantly higher than the bias of other methods, except for Atomic absorption photometry

cantly higher bias. The reference method has excellent performances in terms of variability and bias. From the other methods, only Direct potentiometry - ISE - Roche - Integra showed a comparable performance.

4.6 Belgian EQA for semen analysis

4.6.1 Introduction

Semen analysis is a routine analysis for assessing male fertility. Several factors of the sperm are assessed, of which the concentration of sperm cells is one. It is performed by counting the number of cells in a counting chamber under controlled conditions. It is a manual process and has suffered from

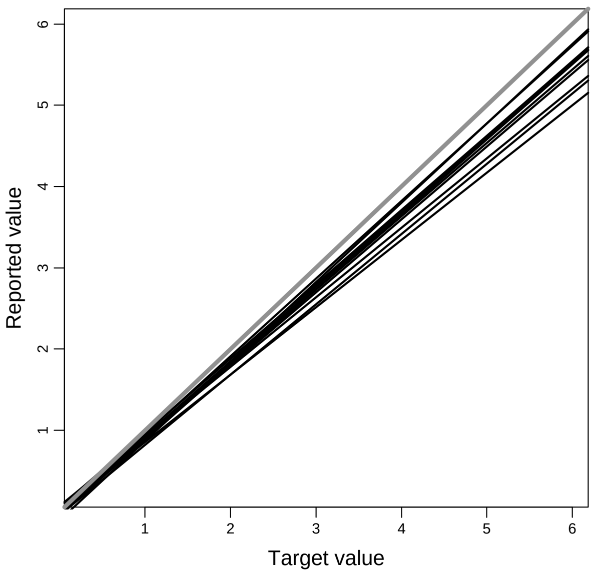

Figure 4.4 Regression lines of the “Thermo electron” method, used by 11 laboratories in the 2010 Belgian EQA survey for Lithium. The thick grey line is the 45° line.

major variability in the past. For this reason, the Belgian EQA for semen analysis has aimed at contributing standardization and quality assessment. The programme started in 2004 and data have been registered in a uniform way since then end of 2005. The 3-step method has been applied to the data obtained since then. In total, 172 different laboratories have participated in the study period, of which 105 have participated during the whole period.

Count data have a tendency to be asymmetrically distributed, as illustrated in Fig 4.5. For this reason, target and reported values have been log-transformed before analysis. The analysis was subsequently performed in the same way as the other analyzes in this chapter. The first period ranged from 2005 to 2007, the second period from 2008 to 2011.

4.6.2 Results

In total, 1633 results obtained for 14 samples were reported from 2005 to 2007 and 1931 results obtained for 16 samples were reported from 2008 to

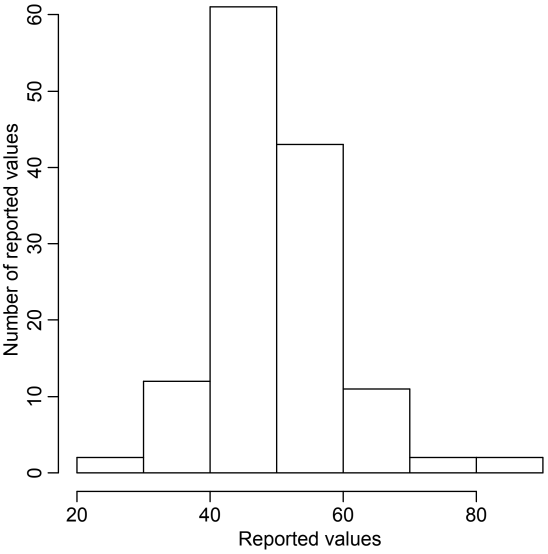

Figure 4.5 Typical histogram of reported values for sperm cell counts (sample SP/1101-1 of survey 2011/1) for laboratories using the same method (Improved Neubauer with positive displacement).

2011. The applied methods in the first and second part of the study period are listed in Table 4.13.

The Improved Neubauer method with a positive displacement has become much more popular over time. It is nowadays used by more than 60% of the laboratories. Some other methods have undergone a decline in use, such as the Bürker and Improved Neubauer without positive displacement.

Outliers against the regression line

In the first period, 5.8 % outliers were found, while in the second period, only 2.7 % outliers were recorded (see Table 4.14). The same also holds for the laboratories that have not changed method during the study period: 5.7 with respect to 2.5 The distribution of number of regression outliers per laboratory, for all laboratories, is shown in Figure 4.6.

The evolution was assessed for laboratories that used the same method throughout the study period. The total frequency, merged over all methods,

Table 4.13 Distribution of number of laboratories for each applied methodology in the two study periods for the semen analysis EQA follow-up study, together with the number of laboratories that have applied the same methodology during the two periods. Period 1: 2005-2007, Period 2: 2008-2011.

Method	Period 1	Period 2	Continuously
Disposable chamber	7	14	1
Bürker	22	10	8
Fuchs-Rosenthal (reusable)	8	4	4
Improved Neubauer	26	4	0
Improved Neubauer, Positive disp.	46	84	34
Makler	8	7	4
Microscope slide	3	2	2
Thoma	9	7	3
Total	129	132	56

Table 4.14 Frequency of outliers from the linear regression line for the semen analysis EQA follow-up study. Period 1: 2005-2007, Period 2: 2008-2011.

	Period 1		Period 2		
Method	Total	Outliers (%)	Total	Outliers (%)	
Disposable chamber	14	0	16	0	
Bürker	109	6.42	126	3.97	
Fuchs-Rosenthal	56	8.93	64	7.81	
Improved Neubauer, Positive disp.	500	4.60	574	1.74	*
Makler	56	14.3	64	1.56	*
Microscope slide	14	7.13	16	0	
Thoma	42	2.38	48	4.17	
Total	791	5.69	908	2.53	*

*Outlier rates significantly different between period 1 and period 2

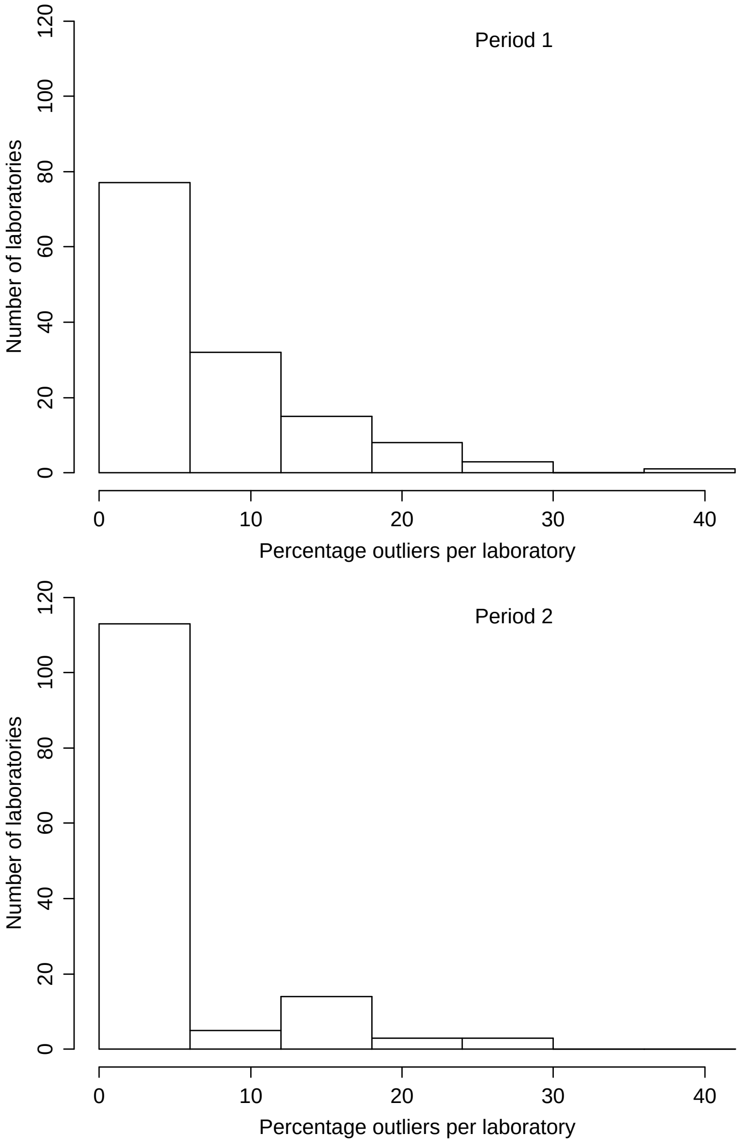

Figure 4.6 Number of outliers against the linear regression model per laboratory for the two distinct periods for the semen analysis EQA follow-up study. Period 1: 2005-2007, Period 2: 2008-2011.

dropped significantly in the second period. It is most clear for the Improved Neubauer with positive displacement and Makler chambers, for which the drop in outlier frequencies was significant.

Variability around the regression line

The results are shown in Table 4.15. A Fisher's exact test for assessing difference in variability outliers between the first and the second period led to a P-value of 0.16, meaning that the observed drop in variability outliers was not significant.

Remark that also a few laboratories that used the reference method, Improved Neubauer with positive displacement, were flagged for exceeding variability.

There was no significant interaction between the period and applied methodology for the regression residual standard error. Neither was there a significant difference between the periods.

For the method on the contrary, a significant difference was found: the residual standard error for the Fuchs-Rosenthal method was significantly higher than for the Bürker and Improved Neubauer with positive displacement.

Bias of the regression line

The number of laboratories flagged for exceeding bias was very low and didn't evolve significantly over time. The Fisher exact test for the difference between the frequency of bias outliers, for all methods together, was 0.2431.

For the absolute bias, there was no interaction between the periods and methods and no significant difference between the periods. For the methods, however, a significant difference was found: the Fuchs-Rosenthal and Makler methods showed a higher absolute bias than the Improved Neubauer

Table 4.15 Count of regression lines with exceeding variability and regression residual standard error for the most popular methods in the semen analysis EQA follow-up study. Period 1: 2005-2007, Period 2: 2008-2011.

Method	No. of Laboratories	Variability outliers (%)		Residual standard error	
		Period 1	Period 2	Period 1	Period 2
Disposable chamber	1	0	0		
Bürker	7	0	0	0.129	0.112
Fuchs-Rosenthal	4	33.3	0	0.183	0.181
Improved Neubauer, Positive disp.	36	5.56	2.78	0.138	0.118
Makler	4	75	25	0.204	0.162
Microscope slide	1	100	0		
Thoma	3	0	0		
Total	55	12.7	3.63		

Note: Fuchs-Rosenthal exhibits a residual standard error that is signifantly higher than the residual standard error from the other methods, except for Makler

Table 4.16 Count of regression lines with exceeding bias and absolute and relative bias for the data obtained for semen analysis EQA follow-up study. Period 1: 2005-2007, Period 2: 2008-2011.

Method	No. of laboratories	Lines with exceeding bias (%)		Absolute bias		Relative bias	
		Period		Period		Period	
		1	2	1	2	1	2
Disposable chamber	1	0	0				
Bürker	7	0	0	0.175	0.127	0.069	0.043
Fuchs-Rosenthal	3	0	0	0.304	0.195	0.122	0.086
Improved Neubauer, Positive disp.	36	0	0	0.097	0.129	0.034	0.052
Makler	4	0	50	0.168	0.054	0.068	0.024
Microscope slide	1	0	100				
Thoma	3	0	0				
Total	55	0	5.45				

Note: absolute bias of Improved Neubauer, Positive disp. is significantly lower than the absolute bias of the other methods, except for Bürker

with positive displacement. The relative bias showed a significant interaction effect between the time and methods. In the first period, the difference between the methods were similar to the differences for absolute bias. For the second period, no significant differences between methods were found.

4.6.3 Discussion

An evolution in accidental mistakes, identified as outliers against the regression line, has taken place over time. Nowadays there are less accidental mistakes and this is the only evolution seen over time.

The results show that the choice for the Improved Neubauer technique with positive displacement as the reference method was optimal: the method has the lowest variability. From the other methods, the Bürker method had the best performance.

4.7 Conclusion

The examples that illustrate the application of the 3-step method have shown that the method is easily applicable to various types of surveys and that it can be easily adapted to specific characteristics of a particular survey, for example in case the regression model suffers from heteroscedasticity. In addition, the link between poor performance according to the 3-step method and mistakes in the analytical process has been investigated and confirmed for two examples for lymphocyte subset counting.

In addition, the method has shown its usefulness for one survey incorporating a relatively large amount of samples and for a historical series of surveys with a small number of samples. Of course, the obtained variability estimate is different: in the first case, the variability estimate is linked with repeatability, while in the second case it is more linked to reproducibility.

Long-term performance of laboratories can be applied in different ways. It may be performed for a cohort of laboratories continuously using the same analytical method over a long period and, as such, yield information about the performance evolution of a certain method. If however the group of laboratories using the same method changes, there is a risk that the evolution of performance over time is biased. Moreover, an improvement over time of accidental mistake rate may reflect a different approach of the EQA organizer, as has been the case for the alcohol surveys. This finding proofs that post-analytical aspects are also in EQA surveys an often overlooked aspect.

The largest prerequisite for application of the method is the setting of reliable target values. Preferably, a reference method, like for alcohol, semen analysis or lithium, is preferred. In absence of a reference method, however, just as for lymphocyte subset counting, the median value can easily be taken as the target value as well. Care has to be taken however and it is adviseable, as done for this particular survey, to verify the technique of the median value with the results of a group of expert laboratories.

Finally, some general remarks can be made that apply to all data analysed in this chapter. An evolution over time was mainly due to a change in frequency of outliers against the linear regression lines and no evolution over time for variability and/or bias was detected for laboratories using the same method over the whole period. These observations may lead to an important conclusion: the evolution of laboratory performance over time is mainly attributable by the decrease of spurious results. The results also show that method performance is a stable factor over time. Hence, wherever EQA organizers observe an evolution of laboratory variability and/or bias over time, a distinction should be between changes in accidental mistakes and pure bias and variability. For the latter, a further distinction needs to be made between an increased use of better performing methods and an increased performance of individual methods. The EQA for sperm analysis, for example, has shown that the best performing method has become more popular in the latest years. In this case, a possible increase in performance in the Belgian laboratories would be due to an increased popularity of the best performing method, rather than to an increased performance of the latter.

CHAPTER 5

Graphical representations of EQA data

5.1 Introduction

The first chapter of this work mentioned the educational role of EQA programmes, in which a non-punitive evaluation is given to each individual participant. Comparing the result of a laboratory with the results obtained by other participants, or by performance goals defined by the EQA organizer, is an important evaluation technique and is greatly supported by the use of graphical representations. Even more, two international standards require the use of graphical representations. The IUPAC protocol for Proficiency Testing [178] for example, clearly stipulates that the EQA organizer should provide to each participant a report with the evaluation of its performance. The report should show the distribution of the results from all laboratories together with the individual score of the laboratory under interest. The results of all laboratories should be provided in graphical form and the IUPAC guide clearly suggests histograms or other distribution plots. Also the ISO

13528 standard [60] requires that EQA participants dispose of tools to interpret their results graphically and describes in detail several options for a graphical representation of the EQA results, mainly under the form of z-scores. In analogy with the IUPAC protocol, it requires that participants dispose of tools to evaluate their results graphically.

Graphical representations should attempt to summarize data by giving as much information as possible, with as few lines, shapes or colors. For representing EQA data, we can add some specific requirements: the quality of the representations should not be influenced by extremely small or large sample sizes and be robust against the presence of a small fraction of strongly deviating data.

This chapter deals with an in depth discussion of several techniques to represent EQA data in participant's reports and looks for further perspectives to make reports more informative for laboratories.

5.2 Graphics for one-dimensional data

5.2.1 Normal quantile plot

Although called Normal probability plots, the Normal quantile plots for displaying univariate series of data are promoted by the ISO 13528 standard. They are constructed by using the quantile function as a link between the original data set and the standard Normal distribution. First, the inverse quantile function is applied to each value of a data series and the result is used to calculate the corresponding quantile of a Normal distribution. Subsequently, the original values are plotted along the Y-axis against the values obtained from the standard Normal distribution along the X-axis.

Normal quantile plots yield information about the shape of the distribution, as well as identification of exceeding points. In the ideal case, when the data are normally distributed and no exceeding points exist, the points in the scatter plot follow a straight line. In case exceeding points exist, the points will be situated around a straight line from which the exceeding points deviate like in Figure 5.1.

Any other deviation from Normality results in a non-linear shape, as can be seen in Figure 5.2. The ISO 13528 standard suggests identifying the points with high z-scores individually. The Normal quantile plot fits for schemes with few and with a lot of participants. For the former, it should be noted that the normal quantile plot has a higher variability and that the tails of the line may deviate from the straight line, even when the data do follow a Normal distribution. In addition, EQA organizers can use the normal quantile plot as a first step to assess the distribution of the reported data and take necessary actions, like transforming data in case of skewness or splitting peer groups in case of multimodality [87].

5.2.2 Histograms

Histograms are probably the best known and most frequently used graphical representation of univariate data series. The range of the data series is partitioned into non-overlapping intervals of equal width and for each interval bars are graphed of which the height is proportional to the number of data inside the corresponding interval. The histogram may be extended with lines representing action limits, for example at ± 2s and ± 3s from the mean value. The ISO 13528 standard suggests to draw histograms of the z-scores. In this way the x-axis is standardized with usually a fixed range and lines at Z =± 2 and ± 3 can be drawn to represent the proficiency assessment criterion.

Although algorithms exist to calculate the optimal interval width, there is ample choice of how intervals are set. The latter may result in histograms showing a different shape, although they are made from the same data (see Figure 5.3). The position of the line of an individual participant informs about the position of the result of the laboratory with respect to the results of the other participants. Histograms also allow EQA organizers to evaluate

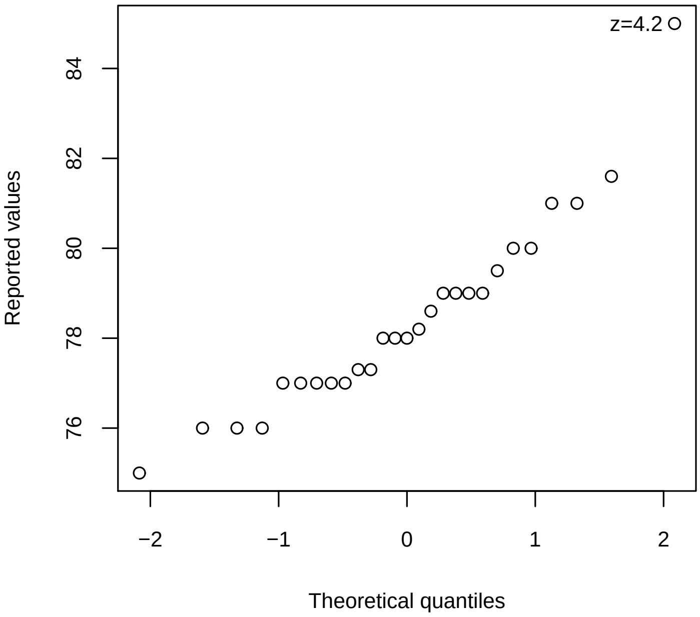

Figure 5.1 Normal quantile plot of reported EQA data for Total protein from one sample by laboratories belonging to the same peer group, with z-scores outside [-3;3] indicated (sample CP/10587 from survey 2010/3, method VIS photometry - Biuret with blank on Cobas Integra (Roche)).

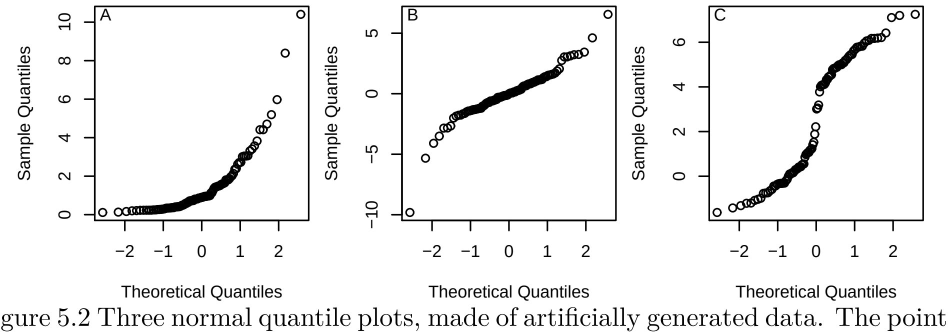

Figure 5.2 Three normal quantile plots, made of artificially generated data. The points in graph A form a line with steadily increasing slope, indicating a skewed distribution. On graph B, the points are situated on a straight line, although the extremes deviate, indicating a leptokurtic distribution. The lines on graph C tend to lie on an S-shaped curve, which is typical for bimodal distributions.

the analytical methodology with respect to performance goals: in the most optimal case, the histograms extend very slightly beyond the limits of the performance goals. In case the limits are too wide and distant from the histogram, performance goals could be made stricter. On the other hand, when the histograms extend too often beyond the limits of the performance goals, the analytical methodology may not meet the required precision and action should be taken as well, either by expanding the limits or by stimulating laboratories to improve their analytical methodology.

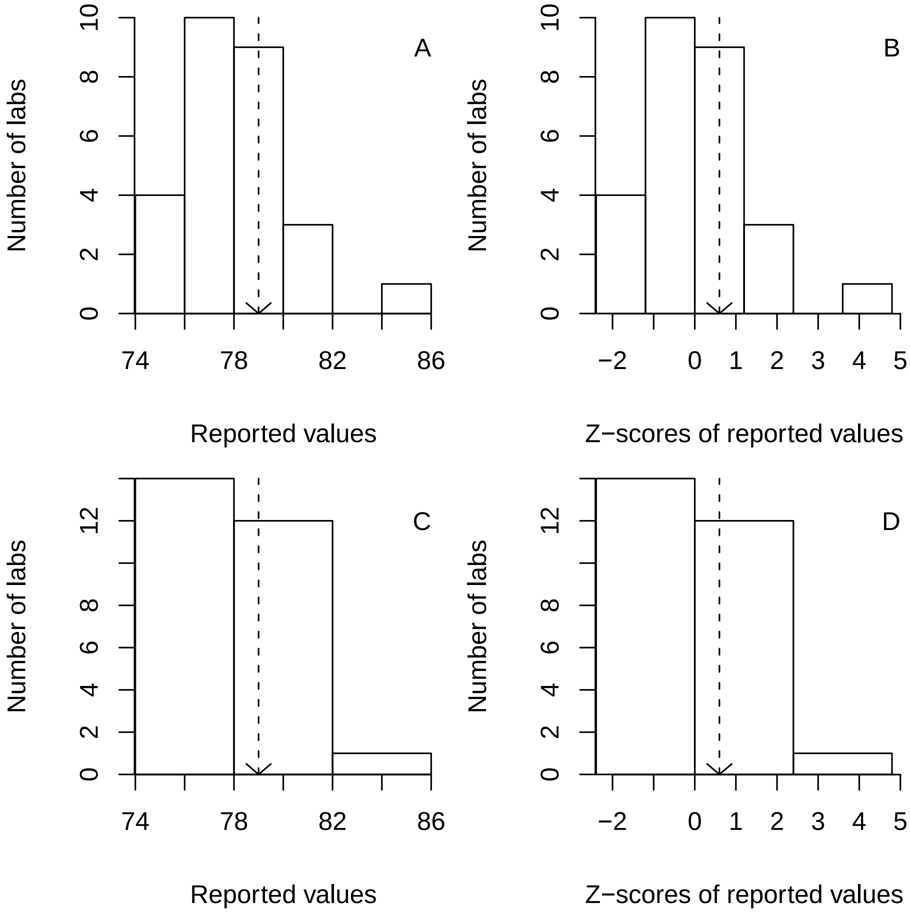

Figure 5.3 Four representations of a histogram of the same data reported for Total protein (sample CP/10587 from survey 2010/3, method VIS photometry - Biuret with blank on Cobas Integra (Roche)). A and C are the histograms made of the reported values, B and D from the derived z-scores. C and D are built with a different choice of intervals than A and B. The black dashed line helps interpreting an individual result.

In case of large outliers, the X axis could be restricted such that extreme observations are not represented on the graph. In this case, a brief indication should be added along the X-axis. Histograms can easily represent a large amount of data, but are less effective for representing data from smaller samples. In the later case, the number of data in each interval may be too low to yield reliable bar heights.

5.2.3 Empirical Cumulative Distribution Function

In contrast with histograms, that show the density of data around a certain value, cumulative distribution functions show the proportion of data smaller than or equal to a certain value. Starting from a sample $\{x_1,x_2,...,x_n\}$, the function $F_n(t)$ counts the number of data x_i that are smaller than or equal to t:

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n}[x_i \leq t]$$

where $[x_i \leq t]$ has the value 1 when $x \leq t$ and 0 otherwise.

The graph is made by plotting $F_n(t)$ against each corresponding value of the ordered data series, as shown in Figure 5.4.

Empirical Cumulative Distribution Functions (ECDF) can be drawn for any $n \geq 1$. They should be preferred to histograms for small sample sizes. Whenever large outliers exist, the X-axis should be truncated to show the majority of the data and a brief indication added along the X-axis. Cumulative distribution functions can easily inform individual laboratories about their position within the group of reported values: for every value along the horizontal axis, the corresponding value along the X-axis shows the number of values not larger than the corresponding value. Hence, these graphs are not only useful for representing reported data, they can also be applied for the visual representation of any kind of performance indicator, as discussed in Chapter 3.

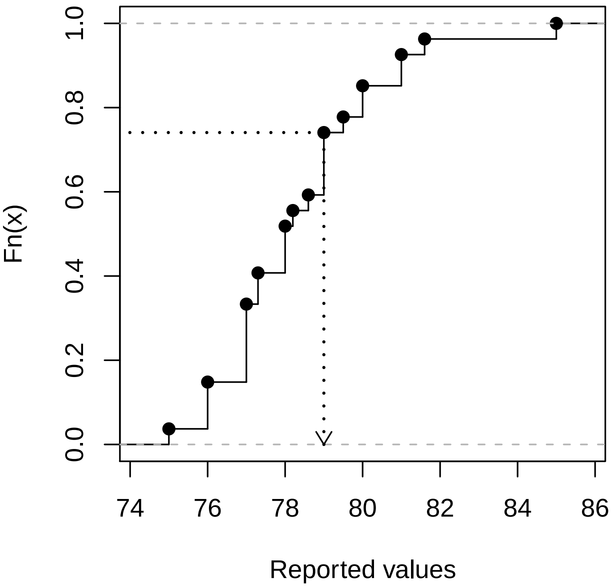

Figure 5.4 An empirical cumulative distribution function for data reported for Total protein (sample CP/10587 from survey 2010/3, method VIS photometry - Biuret with blank on Cobas Integra (Roche)). The dotted line helps interpreting an individual result, showing that there are about 75% laboratories that have reported a value lower than or equal to the individual result.

5.2.4 Box and whisker plots

Box plots are a useful and easy to draw visualization of a univariate series of data. They are made up of the 25th, 50th and 75th percentile (P_{25}, P_{50},P_{75}). A rectangle is drawn from P_{25} to P_{75}. A line at P_{50} is drawn inside the rectangle. Subsequently, the following simple formula is used to identify outliers: a result x is outlying if $x > P_{75}+1.5(IQR)$ or $x < P_{25}-1.5(IQR)$, where $IQR=P_{75}-P_{25}$ is the interquartile range. The thresholds $P_{25}-1.5(IQR)$ and $P_{75}+1.5(IQR)$ are named lower and upper inner fences and in case of a Normal distribution distribution correspond to a z-score of ±2.7. Lines, also called whiskers, extend from P_{25} and P_{75} towards the most extreme values inside the inner fences. As such, they represent the range of all the points, not considered as outliers. Points outside this range are considered as outliers and represented by small dots. In addition, box plots may be used to position the individual value of a certain laboratory with respect to the other values by plotting it as under the form of an extra symbol in or around the box plot.

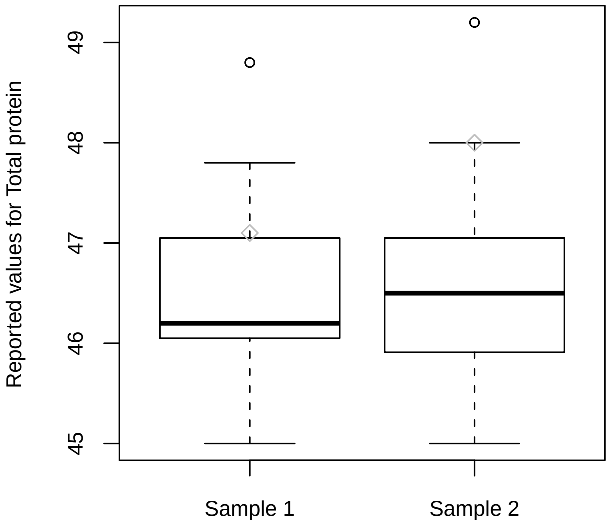

Figure 5.5 Box plots obtained for two samples reported for Total protein (sample C/9947 and C/9948 from survey 2010/1, method VIS photometry - Biuret without blank, Beckman) with similar concentration by laboratories using the same peer group. The grey squares indicate the individual values obtained by a certain laboratory.

Remark that the formula for detecting outliers are a simple and not watertight rule. Other rules exist, for example using limits at $P_{25}-3IQR$ and $P_{75}+3IQR$, also called inner and upper outer fences. Users of software that draws box plots automatically are recommended to check the manual for information about the way how box plots are created.

Just as for histograms, box plots are sample size-independent and can be easily adapted to the presence of large outliers by adjusting the axis and adding a small symbol to indicate that values lie outside of the graph. Even more, they are superior for representing different data series in one plot, because they can give a similar impression about the data distribution with less lines than histograms (see Figure 5.5) .

5.3 Graphs for combining variability and bias

Although not usually done in the clinical setting, EQA organizers may ask the participants to report the uncertainty of the analytical result for a certain parameter. Uncertainties may be calculated in several ways. One may identify all sources of laboratory error, which can be summed up to an uncertainty measure for a certain result. Other measures of uncertainty may be found by calculating a standard deviation on repeated measurements on the same sample, as is often done for analytical method validation.

5.3.1 Normal quantile plot with added variability

A Normal quantile plot can be extended by drawing vertical lines representing the expanded uncertainty around each plotted value [60]. The expanded uncertainty is usually taken as twice the uncertainty, expressed as a standard deviation. Usually, the line is then drawn from the (reported value - 2 × the standard deviation of the reported result) to (reported value + 2 × the standard deviation of the reported result). In addition, horizontal lines may be drawn to represent the expanded uncertainty around the assigned value.

Figure 5.6 shows the resulting normal quantile plot. For correctly estimated variability measures and obtained analytical results, the vertical lines should intersect the horizontal boundary representing the uncertainty around the assigned value.

Care should be taken however when interpreting the results, since the intervals representing the expanded uncertainty could be seen as confidence intervals and two estimates may be significantly different, even when their confidence intervals overlap, as the following example illustrates.

Assume that there are two data series. Let us denote by $\bar{\mathrm{x}}_\mathrm{i}$ and SE_i the

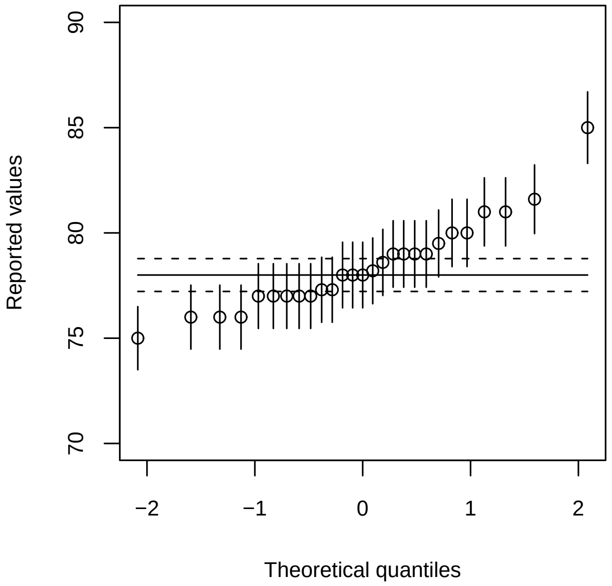

Figure 5.6 Normal quantile plot from reported data for Total protein (sample CP/10587 from survey 2010/3, method VIS photometry - Biuret with blank on Cobas Integra (Roche)) with extended variabilities, that were simulated. Laboratories of which the extended uncertainty interval is completely outside of the extended uncertainty interval of the median have a significant bias.

average and standard error of the mean of group i (i=1 to 2) and that the standard deviations are known beforehand.

A significant difference is found when

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2}} > \mathrm{Q_Z}(0.975)$$

where $\mathrm{Q}_Z(0.975)$ is the upper 2.5.th percentile of the standard Normal distribution. Let us use 2 as an approximate value. Rewriting gives

$$(\bar{x}_1 - \bar{x}_2)^2 > 4(\mathrm{SE}_1^2 + \mathrm{SE}_2^2),$$

On the other hand, two confidence intervals overlap when

$\bar{x}_1 + 2SE_1 < \bar{x}_2 - 2SE_2$ or $\bar{x}_1 - 2SE_1 > \bar{x}_2 + 2SE_2$

This can be rewritten as

$|\bar{x}_1 - \bar{x}_2| > 2(SE_1 + SE_2)$

or

$(\bar{x}_1 - \bar{x}_2)^2 > 4(SE_1 + SE_2)^2$

Remark that $(SE_1 + SE_2)^2$ is larger than $(SE_1^2 + SE_2^2)$ and hence, the test relying on the hypothesis test will find significance for smaller differences between $\bar{x}_1$ and $\bar{x}_2$ than the test relying on confidence intervals. As a consequence, the test displayed in Figure 5.6, as suggested by ISO [60], is non-optimal due to a lack of power.

Other representations that are limited to the special case when uncertainties have been reported and that are not extensions of known graphics, have been reported in literature [48, 157]. They are not discussed in detail here because they rely heavily on reported uncertainties, a practice that is often lacking in EQAs in clinical settings.

5.4 Combining results of different samples

Just as they can give a fast visual impression of a single series of data, graphical representations can be used for representing results obtained from different samples, possibly from different surveys. They are a useful addition to the graphs described before, in which the evaluation was done for each sample separately.

5.4.1 Bar plots of standardized laboratory biases

The ISO 13528 standard describes bar plots in order to give a visual impression of the overall performance of the laboratory, in terms of bias or variability. The results of the laboratories are represented by vertical lines that extend from 0 towards the individual z-score obtained for a certain parameter of a certain sample. Lines representing z-scores obtained from different sample are grouped together and the group of lines for each laboratory are joined in one graph (see Figure 5.7).

Laboratories with high variability are identified by lines that exceed often the ± 2s, or ± 3s limits, in the upper and lower direction, while laboratories with high bias are identified by lines that point frequently into the same direction. Hence, bar plots are able to represent different kinds of deviation from optimality in one graph.

In our opinion, care should be taken however during the interpretation of the bar plots, in particular when only a few samples are plotted. When the laboratory suffers from an increased variability, there is a chance that several lines exceeding z-limits will point in the same direction and hence, suggest bias instead of variability. This chance decreases however with an increasing number of samples taken into account in the graph.

In addition, this kind of representation becomes difficult to interpret for large groups of participants or samples. The vertical axis however could be limited in case of large outliers. In that case, an extra indication can be added to the graph.

5.4.2 Youden plot

The Youden plot is an informative graph to represent data obtained for data from two samples. It is based on a scatter plot, with the results obtained for two samples by the same laboratories plotted against each other. This yields usually an oval-shaped scattered cloud of points.

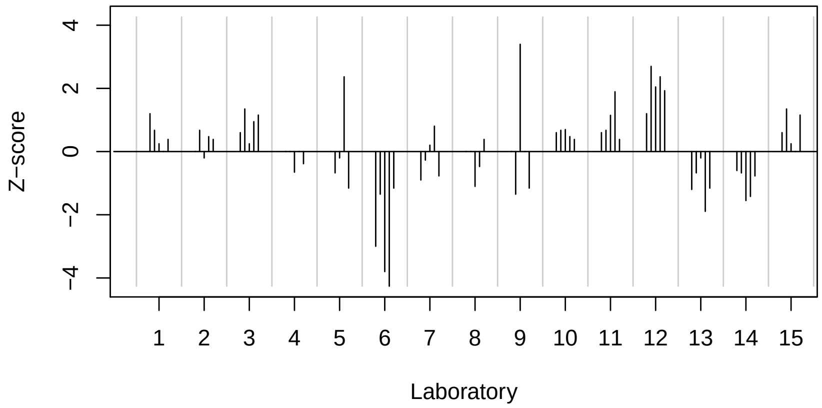

Figure 5.7 Bar plots of one alcohol survey (2010/2) for laboratories that used Headspace chromatography (capillary-column). Remark that laboratory 6 has all its lines far beneath the zero line; the laboratory may suffer from exceeding bias. Laboratory 9 has exceeding lines of which one points upwards and one points downwards. It may have exceeding variability.

A confidence ellipse can be obtained by using the Hotelling T^2 statistic. The ISO 13528 standard notes that the use of the Hotelling T^2 is not robust and hence, strongly influenced by a small fraction of outlying data. Surprisingly enough, the standard mentions that the details of such a method has not yet been worked out, althoug it was published 3 years after the appearance in scientific literature of a robust Hotelling T^2 test [189].

Therefore, we would advise to use the robust algorithm instead of the algorithm mentioned in the ISO 13528 standard.

Depending on the position of the points on the scatter plot and with respect to the confidence ellipse, different zones can be identified and linked to different deviations from the ideal analytical process. Points outside the ellipse and along its major axis may indicate laboratory bias (zones A and B in Figure 5.8), points outside the ellipse and away from its major axis indicate high

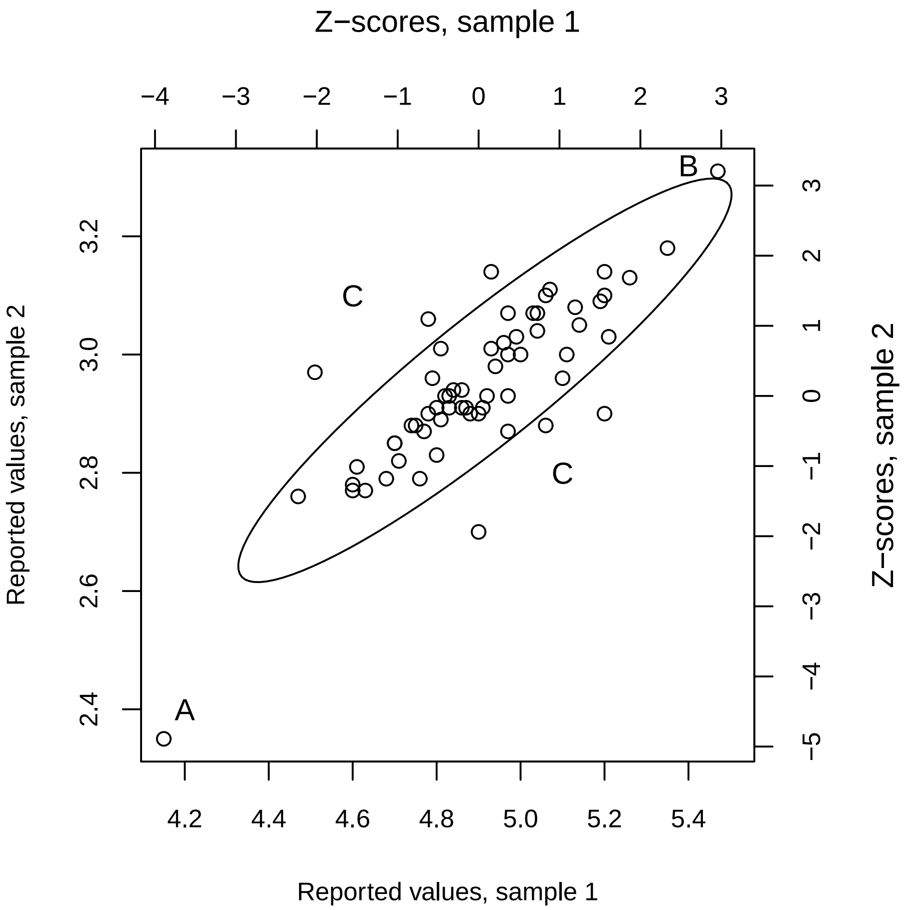

Figure 5.8 A Youden plot of reported values of Thyroid Stimulating Hormone of two samples (T/8070 2 T/8071, survey 2007/4) by laboratories using method "Non-Isotopic Roche-Elecsys / Mod E / Cobas e", with a robust confidence ellipse. Points in Zone A and B are distant from the rest of the data along the major axis and point to laboratories not following the test method correctly. Points near zone C have been produced by laboratories possibly suffering a large analytical variation.

analytical variability (Zone C). Note that the same remark holds as for the bar plots: high analytical variability may also result in points in zone A or B.

EQA organizers may use Youden plots for reporting to participants and for visually assessing homogeneity of peer groups as well. In case the peer group consists of different subgroups with different means, the points will be scattered in different clouds that are distant from each other.

5.4.3 Plots of repeatability standard deviations

The results of multiple measurements of the same parameter on the same sample may be used to obtain a measure of within-laboratory variability. A graph can be made to evaluate the possible bias and variability of laboratories by plotting the within-laboratory standard deviation for each laboratory against the corresponding average value. In the absence of dependency of measurements, a confidence region may be calculated by the following curve:

$$\mathrm{B(x) = s_p exp}\left\{ \pm \frac{\sqrt{\mathrm{Q}_{\chi^2}(1-\alpha,;2) - \mathrm{n}\left(\frac{\mathrm{x}-\bar{\mathrm{x}}}{\mathrm{s_p}}\right)^2}}{\sqrt{2(\mathrm{n}-1)}} \right\}$$

where B is the boundary of the confidence region, $\mathrm{Q}_{\chi^2}(1-\alpha;2)$ is the $(1-\alpha)-$quantile of a Chi-square distribution with 2 degrees of freedom, $1-\alpha$ is a chosen significance level, n is the number of laboratories, $\bar{\mathrm{x}}$ is the grand average of all results, $\mathrm{s_p}$ is a standard deviation obtained by pooling all the individual standard deviations. See the ISO standard for a robust algorithm for pooling standard deviations. $\mathrm{x_i}$ is a continuous variable, ranging from

$$\bar{\mathrm{x}} - \mathrm{s_p}\sqrt{\frac{\mathrm{Q}_{\chi^2}(1-\alpha;2)}{\mathrm{n}}} \text{ to } \bar{\mathrm{x}} + \mathrm{s_p}\sqrt{\frac{\mathrm{Q}_{\chi^2}(1-\alpha;2)}{\mathrm{n}}}$$

The plot is shown in Figure 5.9 for a survey of alcohol determination. Remark that the main axis of the shape of the cloud of points is orthogonal to the main axis of the shape of the confidence regions, meaning that a considerable between-laboratory variation exists, probably due to a small laboratory-specific bias.

This representation is remarkably simple and helps not only identifying laboratories with high bias or variability, it also helps to evaluate the method's ease to reproduce results between laboratories. The method is still reliable in the presence of outlying observations and is applicable to groups of small and large size.

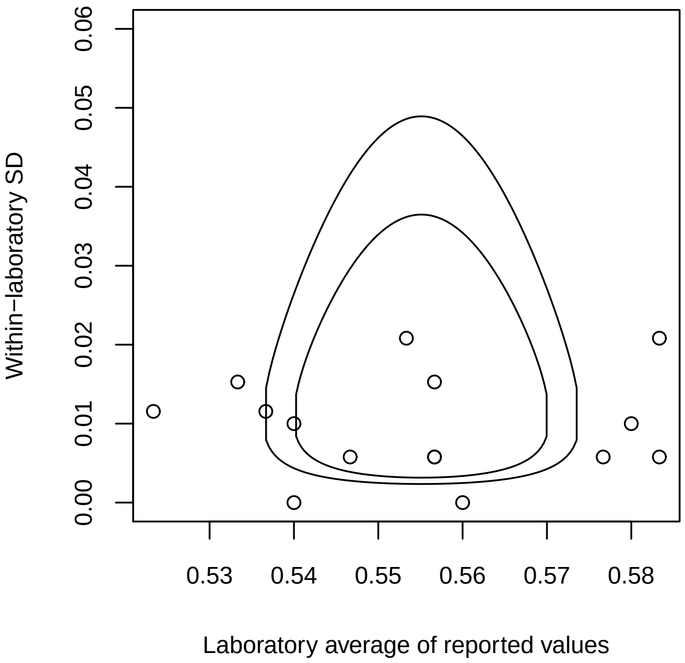

Figure 5.9 Plot of Standard deviation against average for the 2010/1 alcohol EQA survey (samples E/10385, E/10387 and E/10388), reported by laboratories that used Headspace chromatography (capillary-column), with confidence regions at 95% and 99%.

5.4.4 Shewhart control chart for z-scores

Shewhart charts are well known from internal quality follow up and can be applied to external quality assessment as well. The graph can be made by plotting the z-scores obtained for different samples against time. Horizontal lines at ± 1s, ± 2s and ± 3s help to interpret the results. "Out-of-control" points are found beyond the ± 3s-limits. Other criteria for detecting deviant processes in the laboratory may be applied as well, such as two out of three successive points falling outside of the ± 2s limits. An example is shown in Figure 5.10 for the reported results for Ferritine from one particular laboratory.

Shewhart control charts are an excellent tool to monitor the performance of a laboratory over a long time. They can be adapted to extremely outlying observations by controlling the limits of the vertical axis and putting a special mark near the upper or lower horizontal boundary whenever values are situated outside of the graph. There are however some drawbacks. When different samples are sent in one survey, one has to make either a dot plot, in which the z-scores of the different samples are plotted along a vertical line and a line is drawn through the average value for each survey [60], or make Shewhart charts per level of concentration. Then, the EQA organizer should be aware to send each round samples from different concentration ranges. The latter is not always easy to achieve. Moreover, the laboratory may have changed analytical methods during the time spanned by the horizontal axis. The plot should be limited to the periods for which results have been obtained with the same methodology and this has an adverse effect on the length of the time range for which Shewhart chart can be made.

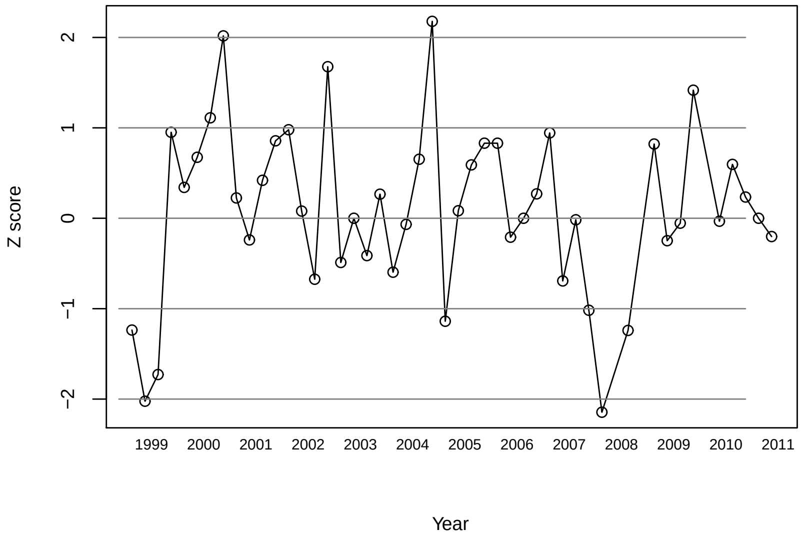

Figure 5.10 Shewhart chart for z-scores obtained for Ferritine, obtained by the same laboratory from 1999 to 2011.

5.4.5 Cusum control charts for z-scores

The cusum control charts for z-scores start from the same model as the Shewhart control charts: the evolution of z-scores over time. A cumulative sum of z-scores is produced over a certain time window, containing several EQA rounds (see Figure 5.11). The cumulative sum is then plotted against time. The graph is mainly an effective method for detecting measurement bias. The graph in Figure 5.11 for example shows that the laboratory may have had a small tendency of a negative bias in 1999 and 2008. The Shewhart z-score chart effectively also shows three consecutive z-scores below -1 for these periods.

Advantages and disadvantages of this graphical representation are similar to those of the Shewhart chart. It may be noted as well that the cumulative sum is always taken over a time window, resulting in a lag for detecting bias that becomes larger with increasing window size. Therefore, considering the relatively low frequency of External Quality Control programmes, it may take long before a bias is detected and hence the effectiveness of this kind of graphs is not always assured.

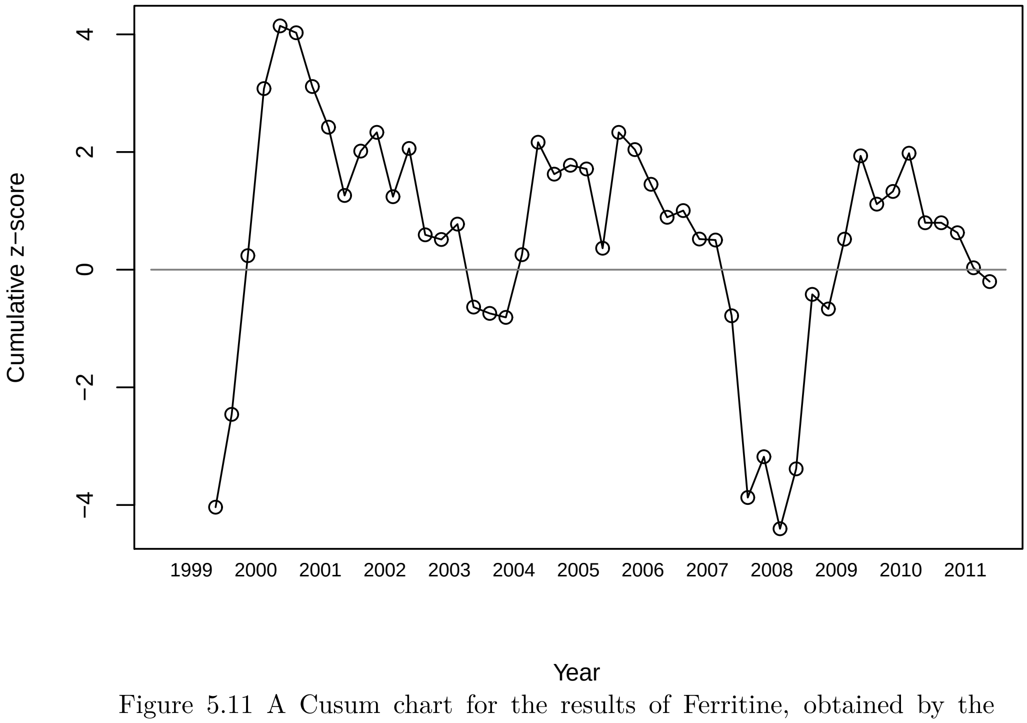

Figure 5.11 A Cusum chart for the results of Ferritine, obtained by the same laboratory from 1999 to 2011. Z-scores were cumulated over three consecutive surveys.

5.4.6 Representation of the 3-step method

The 3-step method, as described in Chapters 3 and 4, is an evaluation technique that helps distinguishing between accidental mistakes, exceeding variability and bias. A visual representation of the method should inform the laboratory about the frequency and nature of its accidental mistakes and give an overview of its variability and bias and, simultaneously, compare the two latter with the results found for other laboratories. Therefore, a graphical representation has been developed consisting of three different graphs: a general representation of the applied regression model, a histogram of variabilities and a scatter plot of intercepts versus slopes, extended with a confidence ellipse.

General representation of the applied regression model

A regression model is easily represented by a scatter plot with the regression line between the non-outlying reported and target values. Specific modifications of the representation here are the addition of the 45°-line and a special indication of outlying regression points, as shown in Figure 5.12. This representation allows a rapid evaluation of measurements reported in a certain time range. The vertical axis can be easily adapted in case of large outliers and also here, a small symbol can be added to indicate that values are situated outside of the graph.

The outliers are supposed to result from accidental mistakes and hence their distribution should be concentration independent. Whenever outliers are more concentrated near the ends of the regression line, however, the linear relation between reported and target values is not assured and laboratories should investigate the departure of linearity in that zone. Comparing the regression line with the 45°-line informs the laboratory about a possible bias of its measurements. In case of a significant bias, a regression line parallel but some distance below or above the 45°-line indicates a constant bias and a regression line cutting the 45°-line indicates a concentration-dependent bias.

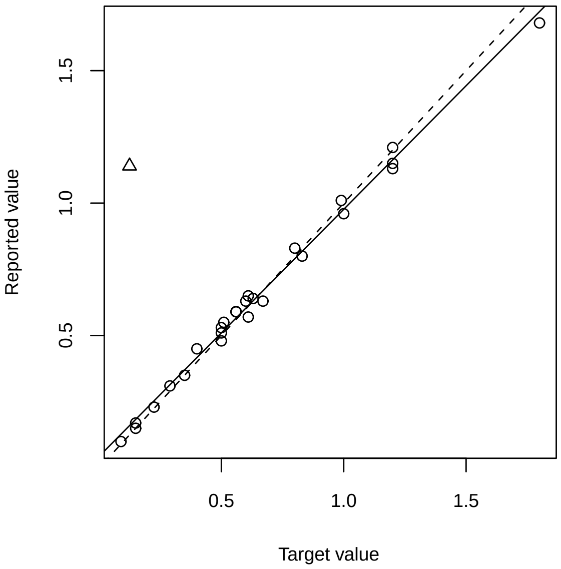

Figure 5.12 The basis of the 3-step method: a linear regression line. The full line is the ordinary least squares lines through all the points except the outlier (indicated by the small triangle), the dashed line is the 45°-line.

ECDF of regression residual standard error

The second step compares the regression residual variability of an individual laboratory with the regression residual variability obtained for other laboratories. An Empirical Cumulative Distribution function, as described before, is one of the recommended ways of displaying a graphical summary of a univariate series of data, in particular performance indicators.(see Figure 5.13). Two extra indications may be added to the ECDF:

(1) lines representing residual variability of the individual laboratory (dashed line in Figure 5.13)

(2) lines representing the upper threshold value for identifying regression lines with exceeding variability (dotted line in Figure 5.13)

The most important comparison to be made here is between the positions of the green and red line. A laboratory is flagged for exceeding variability when the green vertical line (representing the laboratory's individual resid-

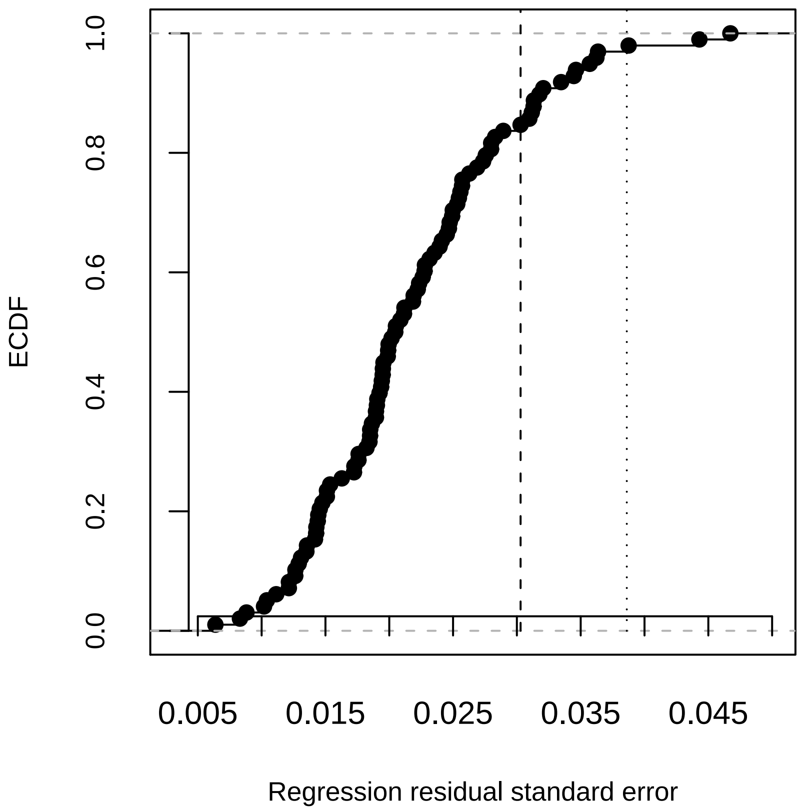

Figure 5.13 ECDF of regression residual standard error values found for a group of laboratories using the Roche enzymatic method in the alcohol EQA survey from survey 2007/1 to 2009/1. The vertical dotted line indicates the upper threshold value for flagging exceeding variability, the vertical dashed line represents the regression residual standard error of a particular laboratory.

ual variability) is to the right of the red vertical line (representing the upper threshold). In addition, the green lines inform the laboratory about its position in terms of residual variability within the whole population of laboratories. The example in Figure 5.13 demonstrates that, although it is not flagged for being exceedingly large, the regression residual variability of this laboratory is relatively high, since more than 80 % of the laboratories have a lower residual variability.

Scatter plot of intercept versus slope

The intercept and slope of a simple linear regression model, with a fixed population residual variance, intercept and slope, are negatively correlated. This negative correlation is still visible when intercept and slope of different variables are plotted against each other. In general, intercept and slope of the regression lines of all laboratories can be considered as a bivariate Normal sample. A robust confidence ellipse using the robust Hotelling T^2 statistic [189] can be calculated around the data made up by the intercepts and slopes and delineate the area out of which regression lines are flagged for exceeding bias (see Figure 5.14).

In addition, the value (0,1) may be drawn on the scatter plot, together with the mean intercept and slope of all the laboratories not having exceeding bias. The analytical method's bias may be evaluated by considering a confidence region around the mean intercept and slope: the method has a significant bias when the (0,1) point is outside of the confidence region around the mean intercept and slope. Note that the latter may loose power when there is a large variability between the position of regression lines of the different laboratories and/or regression lines are calculated over a small number of points.

5.5 Towards interactive plots

Traditionally, reports of EQA surveys are produced on paper or more recently, electronically, such as a Portable Document Format (PDF) file before sending to the participants. The graphical representations discussed so far have already demonstrated their efficacy in providing information about the laboratory performance and can be easily drawn on paper or in a PDF. There is however one major drawback of distributing graphs on paper or in a PDF: they are rigid and lack any kind of interactivity. The kind, form and selection of data to be shown is a choice in the hands of the EQA organizer. As a consequence, the laboratory, sometimes reporting results for many years

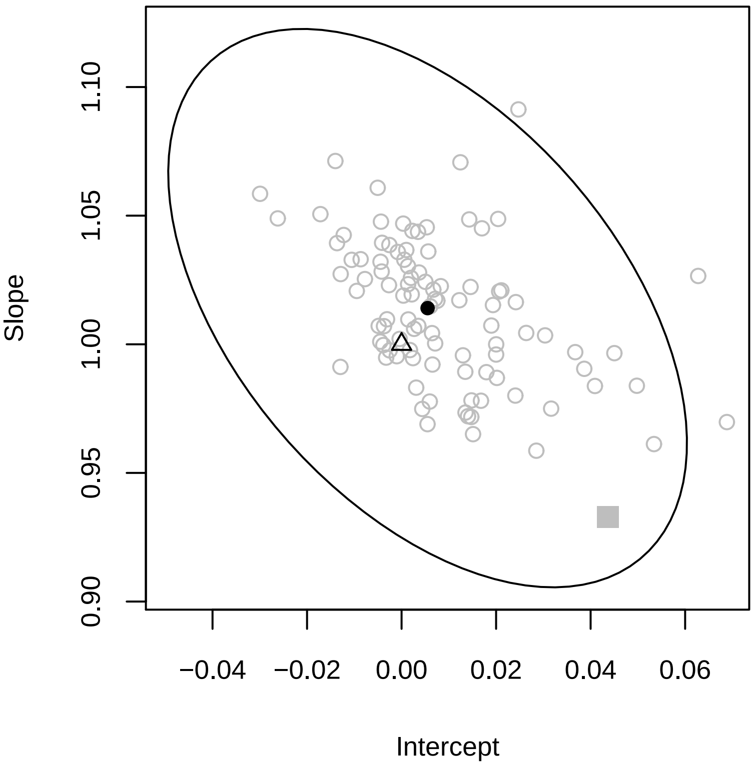

Figure 5.14 Evaluation of the last step of the 3-step method: bias, via a plot showing intercept versus slope for a group of laboratories using the Roche enzymatic method in the alcohol EQA survey from survey 2007/1 to 2009/1. The large grey quadrangle represents the intercept and slope of an individual laboratory. Grey round points represent the individual intercept and slope of other laboratories using the same methodology. The triangle represents the 45°-line (intercept 0, slope 1). The ellipse is the robust confidence region for the (intercept,slope) combinations of individual laboratories.

consecutively, has no way for a further graphical exploration of its results than overviewing several graphs from different reports next to each other.

Interactive graphs on the contrary enable the laboratories to display results from different rounds according to their own choice and may help them understanding their results in a better way. For example, the visualization of the 3-step procedure uses a selection of consecutive surveys as a time frame and a graph on paper follows the EQA's organizer choice of length and position of the time frame. The time frame length may also be important for the

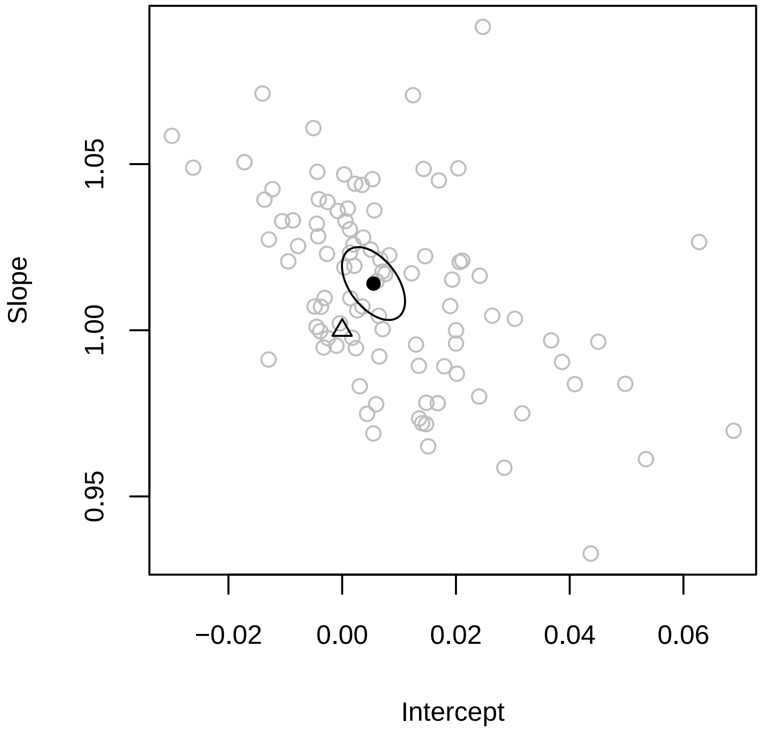

Figure 5.15 The evaluation of a method's mean bias for a group of laboratories using the Roche enzymatic method in the alcohol EQA survey from survey 2007/1 to 2009/1. The grey points are the (intercept,slope) combinations of the individual laboratories. The black dot is the method's mean combination of intercept and slope, the ellipse is its confidence region. The triangle is the point representing the 45°-line. Since the latter is outside the confidence region of the method's mean, the method has a significant bias.

cumulative z-score charts, where a laboratory may choose the time window such that it contains solely results obtained with the same methodology. Of course, graphs on paper or in a PDF file are not suited for this purpose but a solution could be worked out within a web page.

Interaction with databases is supported by several programming languages and applications within web pages are relatively easy to distribute and maintain. A solution should exist of graphics that are able to interact with the user by means of event handlers, mainly triggered after a mouse event. Also, an interaction with the HTML Document Object Model (DOM) is advisable, such that an interaction between vector graphics and other HTML elements can be assured, for example offering text-based information, triggered by an event in the graph. For example, a solution that allows the user to click in a graph on a certain symbol, should trigger an action to show more information about the data behind the symbol in text-based format elsewhere on the page. Three possible technologies may be considered for this purpose: Adobe Flash, Java and Scalable Vector Graphics.

(1) "Shockwave Flash format (SWF)" objects from Adobe Flash are widely used. They are produced and compiled before being put on a web server. With the help of an installed plug-in, they are loaded on the client machine and can be 'played'. SWF files, however, consist of a binary file, which doesn't allow easy text-based searching. They are able to access HTML DOM objects but the reverse is not always possible.

(2) "Java applets". They are small binaries written and compiled in Java that can run in any browser on any hardware or software platform. Also here, a plug in, the Java Virtual Machine (JVM), has to be installed before Java applets can be used. Java is a general-purpose language and has a specific Application Programming Interface (API) for drawing and interacting with 2D graphics. Also, it has an extensive database integration. However, Java is a heavy-weight option, often characterized by slow loading. Just as SWF files, Java applets are binaries that don't allow easy text-based searching and lack an easy interactivity with the HTML DOM. An example of a Java applet for drawing 2D graphs are S-PLUS graphlets. The interaction they have with the HTML DOM is unidirectional: HTML elements can be accessed and modified from within the graphlets, but objects on the graph cannot be accessed from within HTML objects or functions.

(3) "Scalable Vector Graphics (SVG)". They are an open standard developed by the World Web Consortium (W3C) since 1999. They can be briefly described as an Extensible Markup Language (XML) format for describing two-dimensional vector graphics. From the three options,

they are the only W3C recommendation and their XML format allows easy text-search facilities. They have a complete integration with the HTML DOM, so they can be embedded in any web page and interact with any other HTML or function defined in the page. With the use of AJAX technology, an easy integration with databases is possible and the graphics are able to respond quickly to user interactions. Even more, in contrary with the other two suggestions, SVG is an official standard supported by the W3C, which means that it should be supported by the major browsers. Nowadays the most recent versions of all the major browsers support SVG. The following paragraphs will explore the SVG technology in further detail.

5.5.1 Scalable Vector Graphics

Information coded by XML is stored in text format. It should be written in a structured way to be valid. The structure is made up of tags, which are bounded by the '<' and '>' characters. There are three types of tags, begin tags (<..>), end tags (</...>) and tags without elements (<.../>). For the special case of SVG, every part of the graph, for example a line, polygon or piece of text, is written down in a tag. The tag for the structure of a text element for example looks like:

```
<text x="100" y="190">X</text>
```

The first word in the tag defines it type. In this example, the word 'text' indicates a text element. In general, including a text element into an SVG file will put a certain piece of text on a certain position on the graph. Between the word text and the > several attributes can be specified. In the above example, the x and y attribute indicate that the text should be written on the (x,y) coordinate. Note that SVG coordinates are not like the coordinates in a usual graph: while readers may be used to put the origin left beneath the plotting area, SVG puts it in the upper left corner is (0,0). As a consequence, the Y axis of the coordinate system points downwards. The content of the text to be written is, in XML language, called the element and is to be found between the start (<text...>) and end (</text>) tag.

Example of an SVG document

An example of an SVG document can be found hereafter. Note that the line numbers have to be removed before the file can be interpreted by a computer.

```
 1. <svg xmlns="http://www.w3.org/2000/svg" version="1.1"
 2.     width="200" height="200" >
 3.  <g id="axes">
 4.   <line x1="20" x2="180" y1="180" y2="180" id="xaxis" stroke="black"/>
 5.   <line x1="20" x2="20" y1="20" y2="180" id="yaxis" stroke="black"/>
 6.   <text x="100" y="190" >X</text>
 7.   <text x="10" y="100" >Y</text>
 8.  </g>
 9.  <g id="points">
10.   <circle cx="130" cy="130" r="2" stroke="black" fill="white"/>
11.   <circle cx="30" cy="125" r="2" stroke="black" fill="white"/>
12.   <circle cx="56" cy="135" r="2" stroke="black" fill="white"/>
13.   <circle cx="60" cy="85" r="2" stroke="black" fill="white"/>
14.  </g>
15. </svg>
```

The output of the file is visualized in an SVG interpreter as shown in Figure 5.16.

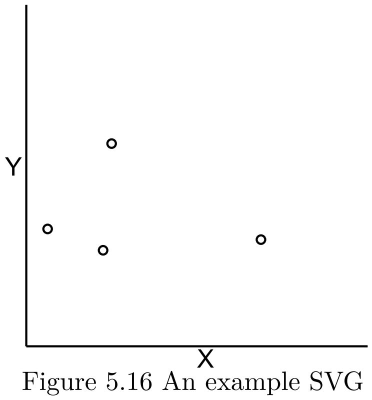

Figure 5.16 An example SVG graph.

An SVG document is in fact an XML file with some specific tags. Every SVG element should start with the <svg> tag. Specifications of the name space, version, width and height are given. Note that any number of white

spaces may be put in between different attributes. In this case, the SVG file will have a width and height of 200 pixels. The 3rt to the 8th line contains the first group of elements. Elements are grouped by the <g...></g> tags. Grouping elements becomes interesting when one wants to add or remove them all together or apply geometrical transformations The 3rd and 4th lines contain the specifications of the horizontal and vertical line. Since they don't have elements specified, they are written down by the <.../> tag. Lines 5 and 6 define the names of the lines as text elements. Lines 9-14 define the groups of points that will be put on the graph. Each point is defined by a separate <circle.../> tag. The center of each circle is defined in the cx and cy specifications and the radius in the r-specification.
This example demonstrates that the specification of attributes is a crucial part of writing SVG. In fact all location, shape, size and appearance specifications can be written down by specific attributes.

An extensive list of specifications can be found at the web page of the W3C (http://www.w3org/Graphics/SVG)

Attributes are also used to describe behavior with respect to mouse events. If, for example, we want the circle defined on line 12 to be blue inside when the mouse moves over it and white otherwise, we can write:

```
<circle cx="60" cy="85" r="20" stroke="black" fill="white"

   onmouseover="this.setAttribute('fill','blue')"
   onmouseout="this.setAttribute('fill','white')"/>
```

The setAttribute function, as shown in the previous example, is a very useful Javascript method to change dynamically attributes of an SVG element. In fact, the function empowers an easy and elegant modification to the location, shape, size or appearance of any SVG object. Moreover, SVG documents can be completely integrated into an HTML page, while SVG elements can be created and their attributes changed on the fly. For example, the following piece of code describes an HTML page that contains an empty SVG element and to which elements can be added, modified, or taken away:

```
1. <html>
2. <script >
3. function drawcircle()
4. {
5.  var D=document.getElementById("ob");
6.  S=D.getSVGDocument();
7.  ln=S.createElementNS("http://www.w3.org/2000/svg",
8.               "circle");
9.  ln.setAttribute("cx","150");
10.  ln.setAttribute("cy","150");
11. ln.setAttribute("r","20");
12. ln.setAttribute("fill","black");
13. ln.addEventListener("mousemove",function(e){
14. this.setAttribute("stroke","red");
15. this.setAttribute("r","100")
16.},false);
17.ln.addEventListener("mouseout",function(e)
18.{
19. this.setAttribute("fill","black");
20. this.setAttribute("r","20")},false);
21. S.documentElement.appendChild(ln);
22.} ,false);}
23.</script>
24.<body>
25. <object id="ob" data='svg.svg'
26.  width="300" height="300" type="image/svg+xml">
27. </object>
28.<input type="button" value="Draw circle"
29.       onclick="drawcircle()">
30. </body>
31. </html>
```

The file svg.svg contains an empty SVG document:

```
<svg xmlns="http://www.w3.org/2000/svg"></svg>
```

It is best to start reading the code on line 25, where an SVG object is defined with a width and height of 300 pixels. On lines 28-29, an HTML element, a button, is defined that will start the function 'drawcircle' when the user clicks on it. The function drawcircle to be found at lines 3 to 22 subsequently paints a black-filled circle on the SVG canvas. The circle becomes larger and red when the user moves the mouse over it (lines 13-16). It becomes again black and smaller when the user moves the mouse away from it (lines 17-22).

The above examples demonstrate the possibilities and flexibility of embedding SVG into HTML: Javascript functions, that enable HTML elements to react on the user's behavior, can also be applied to SVG elements and a complete interaction between SVG, Javascript and HTML elements is possible.

5.5.2 Long term evaluation of EQA results by SVG

Two examples are given to demonstrate the possible use of SVG for reporting EQA data. The examples contain data from the Belgian External Quality Assessment scheme of alcohol and could easily be adapted to other surveys as well. The alcohol surveys of the Belgian EQA scheme are run twice a year. At each survey, 5-6 fresh serum samples are sent. The two applications try to support the pedagogical role of the EQA scheme, in providing a framework to the participants for exploring their own results and to compare them with the results obtained by the other laboratories.

A basic element of the two applications is the time selection element, as shown in Figure 5.17. The surveys are identified by the year during which they were executed and their ranking number within the year and are represented by two lines of text: the upper line represents the years, the lower line the ranking. A selection of surveys is made by positioning a shaded rectangle over them. The position and width of the rectangle is controlled by four small triangles to the right of the two lines of text. Clicking on the left or right triangle makes the grey rectangle to move, respectively, to the left and the right and, as a consequence, another set of surveys is selected. Clicking on the upper or lower triangles makes the rectangle, respectively, to extend or shrink. As a consequence, the four rectangles allow the user to select surveys according to a moving time window, of which the length can be adapted. The two examples given will change the data shown in the graph according to the selected surveys. They can be accessed via http://www.jewidaco.be/chapter5/examples.html.

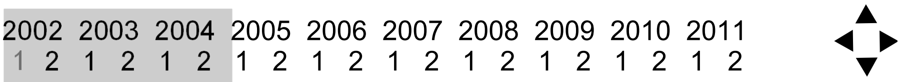

Figure 5.17 Time selection element, used in the two SVG applications. The two text lines represent the surveys, 2 per year. Surveys for which no data were reported, such as 2002-1 in this case, are shown in grey.

An interactive environment to explore z-scores

Traditionally, z-scores are graphically accessed over time by displaying them on Shewhart or Cusum charts (see Figure 5.10 and 5.11), in which they are plotted against time. From the viewpoint of internal quality control procedures, plotting z-scores against time is an excellent way to evaluate the measuring process and the right application of the rules allows a rapid interaction in case the measuring process would be out of control. However, internal and external quality control procedures differ in several aspects. The frequency of measuring control samples in EQA, for example, is much lower and it is not the basic purpose of external quality control to detect rapidly anomalies in the analytical process.

As mentioned before, the education role of EQA programmes is becoming more and more important and in the light of this, it may be important for the laboratory to explore its results with respect to other variables than time, like, for example, concentration. A plot of the z-scores with respect to the assigned values (see Figure 5.18), for example, may reveal information about a certain long-term bias that may be concentration-dependent.
Also, the number of times the z-scores exceed some limits in a particular concentration range may help the laboratory to search for appropriate actions. The combination of this plot with the moving rectangle to select consecutive surveys, adds an extra dimension to the plot: the ability to rapidly

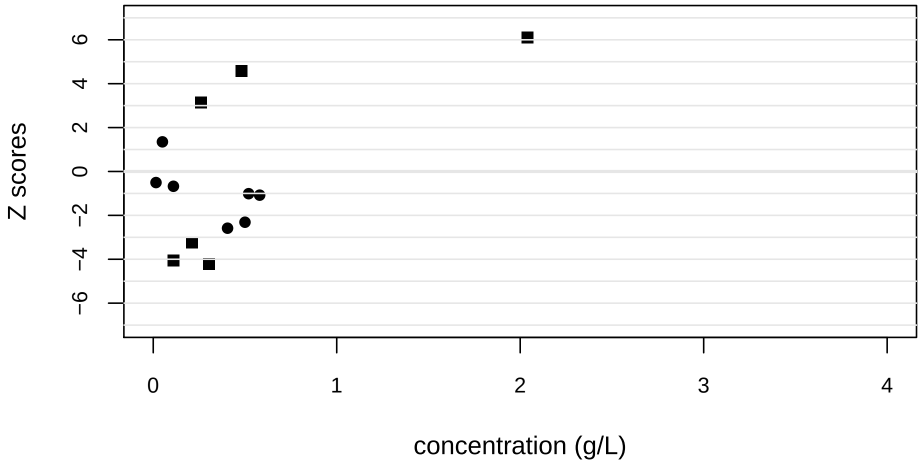

Figure 5.18 An example of a graph of z-scores against concentration, shown for one particular laboratory for selected surveys for alcohol determination. Remark that about the half of the points (squares) exceeds the 3s limits, indicating a poorly performing laboratory.

select different surveys allows to user to evaluate its z-scores with respect to concentration and time simultaneously. In addition, the system reveals extra information about certain z-scores when the user clicks on them. The example that can be accessed from the web page. It has four built-in examples, which can be accessed by clicking on them on the right part of the page.

The z-scores are shown in green when they fit within the ± 3s boundaries and red when they are outside.

The first laboratory has a remarkable high frequency of z-scores outside of the ± 3s boundaries. Remark further that the laboratory has a high frequency of negative z-scores in the beginning and a high frequency of positive z-scores for more recent surveys.

The second laboratory is a very good performer: it has almost all its reported results within the +/- 1s boundary. It is interesting for this laboratory to increase the time frame size such that it covers all the surveys: all the green points are situated randomly around and very closely to the zero line.

For the third example, we can see that the frequency of z-scores beyond ± 3s decreases for the recent surveys.

The fourth example, at last, should be explored by shrinking the size of the time frame to one survey. Although the results rarely fall beyond the ± 3s limits, it is clear that this laboratory has a small negative bias: for several surveys, all z-score values are negative.

An interactive environment to explore the results of the 3-step method

Long-term performance indicators, like the 3-step method described in Chapters 3 and 4 of this work, assume a constant performance within the time range under consideration. For this reason, the time frame for evaluation should be as short as possible. On the other hand, several data are needed to obtain a representative view of the laboratory performance and for this reason, the time frame for evaluation should be large enough.

Whenever EQA organizers issue a report with the evaluation of results obtained in several surveys, a time frame should be selected and unfortunately enough, there is no general rule for calculating an ideal time frame to highlight the performance of all individual laboratories. Sometimes laboratories have recently renewed their methodology and they may prefer a short time frame that contains only data from the surveys for which the new methodology has been used. Other laboratories may want to have an overview over a longer time to check the stable performance of their method.

An application has been developed using the same time selection item described before in which the user can freely select the surveys for which the data will be plotted. The graphs depicting the results of the 3-step method (see Figures 5.12, 5.13 and 5.14) are each time redrawn after a new selection of surveys and in this way, a user-driven selection of surveys can be made. Users are able to select the window such that it spans over a series of surveys for which a homogeneous methodology has been applied. The graphs also allow to give extra, non-graphical information. For example, details about sample and survey name are given when the user clicks on one of the points in the regression graph. In addition, the graph showing the results of the first step has a zoom-in function. By dragging the mouse inside the graph, a rectangle is drawn. Releasing the mouse button rescales the graph such that a zoom on the points inside the rectangle is created. Clicking on 'reset zoom' shows again the full graph.

The first example is a laboratory that suffers from a high frequency of outliers, mainly for concentrations above 1 g/L. The high frequency of outliers can be visualized easily by extending the survey selection rectangle to all the surveys.

The second example is a laboratory showing a high variability. The laboratory is flagged for increased variability for as good as any subset of three continuous surveys, except for data reported from end of 2009 on. This example proves that moving the time selection frame forwards and backwards enables the user to evaluate the evolution of laboratory performance over time.

The last example shows a laboratory that showed exceeding bias in the beginning of the period. However, it has been able to control bias and for the latest surveys, not any deviation was found any more. When the time selection frame is set at a width of four surveys, one can clearly see that the point representing the intercept and slope of the regression line was at the edge of the cloud of points and often situated outside the confidence region. Starting from 2008 the data don't show any bias any more.

5.6 Conclusion

Graphical representations of EQA data are useful tools to give a visual impression of the reported results. They are not only a helpful part of the reports sent to the particapiting laboratories, they also serve to the EQA organizer for quickly overviewing the reported results and searching for discrepancies in the data, like violations of the distributional assumptions. Two types of graphs can be distinguished.

The first type visualizes the distribution of the data and with the help of some additional symbols, individual participants can position themselves with respect to the other participants. They include normal quantile plots, histograms and box plots, of which the two latter are the most popular and can serve as each other alternatives. Histograms are easy to understand and enable an easier positioning of an individual value with respect to the other results; box plots are easier to be drawn and are more suitable to summarize different data series in one graph.

The second type include graphical representations of specific questions. They include the Shewhart type and Cusum charts, graphs that combine bias and variability and the graphical representation of the 3-step method. For the 3-step method, a small comparison of output in Chapter 4 and Chapter 5 illustrates that the graphical representation of results yields a much better understanding of the nature of mistakes. For this reason, the graphical representation should be preferably used whenever the method is applied in an individual report.

EQA data, however, have some specific characteristics that should be taken care of in any kind of graphical representation, since heavy outliers may occur. A simple redesign of one of the axes may help and an indication will be added as soon as a value has fallen outside the graph. In addition, the interpretation of some graphs may be more cumbersome than at first naive sight. When the expanded uncertainty is shown, for example, one should bear in mind that significant differences between two populations are not the equivalent to non-overlapping intervals set up by expanded uncertainties. In addition, in particular for graphs showing bias and variability, care has to be taken with the assumptions of possible dependency between data.

Finally, robustness of specific curves should be an issue and implemented everywhere. For the robust Youden plot, for example, the theoretical elaborations are available.

A major improvement for reporting data to EQA participants, definitely for displaying data obtained over a longer period, is the use of interactive graphics. The technology to provide interactive graphical representations to the participants, under the form of SVG, exists and is well fit for the purpose of providing interactive graphs of EQA data. They enable a framework in which the EQA organizer can decide the content and the form of the data to be shown and the participant has the freedom to explore data in his/her own way, for example by selecting various subsets of surveys, or by displaying textual information near the graphed points of interest. In this way, the EQA participant is supported in the utilisation of exploring its own results in order to find deviations from the ideal laboratory process. Also, the interactive use of graphics may raise the awareness of mistakes that are overseen by EQA organizers and may stimulate participants to explore their results in a much further extent than is nowadays possible.

Discussion and Conclusions

External Quality Assessment Programmes for clinical laboratories have been running for more than half of a century [121]. In Europe, they have mainly evolved towards programmes with an educational purpose as opposed to proficiency testing schemes with punitive sanctions as in the US. Today, External Quality Assessment is considered as a crucial part in the total quality management of clinical laboratories. It is also a vivid scientific discipline with a bunch of articles published monthly in peer-reviewed journals.

External Quality Assessment is facing major challenges; in particular novel statistical approaches need to be developed to help improving the effectiveness and service of EQA programmes [45]. Obviously, EQA programmes should respond to changing needs in the field of laboratory medicine, which has undergone considerable evolution in the last decades. Historically, EQA focussed mainly on the analytical aspects of laboratory work [125]. Recently, interest has shifted to assess pre-analytical and post-analytical phases as well [82, 78]. We believe that major improvements in EQA schemes can be achieved by being able to screen for pre-analytical and post analytical errors. Currently, questionnaires and/or check lists are sent to the laboratories to assess these two critical phases [147, 1, 73]. EQA organizers should try to close the gap between their current working processes and those in the clinical laboratory by mimicking the laboratory processes of the pre-analytical and post-analytical phases as well. For example, specific sample material could be used to assess the behaviour of the laboratory staff with respect to handling uncommon samples or clinical cases [177, 82, 121]. In addition, automatized reporting systems [91] from the laboratory to the EQA organizer should be implemented in order to assess post-analytical mistakes. Assessing mistakes in the pre- and post-analytical phases should be done by modelling their frequency [186]. Several statistical models can be utilized for assessing the analytical phase, where classical z-scores have proved to be an effective and well understood concept to assess laboratory performance. Various algorithms for calculating z-scores have been described [60, 50, 44, 30, 148, 191, 140] and they can be divided into two categories: the first one consists of techniques based on robust statistics, whereas methods of the second category proceed with outlier searching algorithms prior to calculate classical means and standard deviations [57]. For small sample sizes (n less than 20), an approach in which outliers are first removed using Grubbs' test and then mean and SD are calculated on the remaining values has shown to be superior to robust techniques. In presence of outliers, however, Tukey's approach yielded the least biased estimate of variability. Hence, when robust approaches are preferred, we suggest Tukey's approach. In any case, z-scores should not be calculated for peer groups including less than 6 participating laboratories [26].

A second challenge of EQA programmes is to bring standardization among methods [65]. Today, different analytical methods don't necessarily deliver the same test results for a particular sample, although directives, such as the EU Directive on IVD products [36] that states that "methods should show traceability to standards of higher order", are one step towards standardization. In our work, it has been demonstrated that standardization had not been attained. For example, deviations were found in an EQA programme for estradiol and progesterone for various methods [27]. In addition, EQA organizers have to be aware of the quality of the control samples they use. Often, the control material is treated before it is sent to the participants, so that it can be distributed on a large scale and preserved for a longer time period. As a result, matrix effects problems occur, in the sense that differences in test results will appear between methods that could be attributed to the preparation of the sample material and that would not occur for fresh human samples [98]. Not only should EQA organizers strive towards the use of material that is free of matrix effects for a large series of parameters, they should also carefully select the peer groups and restrict themselves to make comparisons between laboratories that use the same or equivalent methodology.

Further, whenever EQA organizers have historical data at their disposal, they should consider the results from different samples as different variables spanning in a multivariate space. Multivariate techniques can be applied to reveal differences between groups of laboratories that may be too large to obtain reliable estimates of standard deviation if the laboratories were gathered into a single peer group. Whenever this occurs, EQA organizers should split up peer groups according to more similar analytical methods. To perform inter-method comparisons, a broad spectrum of possible techniques can be found in literature. Orthogonal regression is usually preferred to simple linear regression when analytical techniques are compared in the same laboratory using various samples [89, 38]. When a target value has been determined with a reference method, however, EQA organizers may compare individual methods with the reference one [175]. Because of the possibility of matrix effects whenever control samples have undergone preparatory steps that routine samples don't undergo, we suggest to check for linearity using a technique that is based on the bootstrap and which can find deviations from linearity, while allowing for a deviation from linearity that is not clinically relevant [27]. Whenever matrix effects can be excluded, for example by using fresh and untreated sample material, and target values for each sample can be determined, we propose another approach which can distinguish different types of deviations from the ideal process of laboratories reporting values equal or close to the target values. This approach consists of three different steps [28]. First, a regression model can be built for each laboratory separately and outliers against the regression model will be identified using robust techniques. These outliers can often be linked to accidental mistakes, which can in turn often be traced back to errors in the pre- or post-analytical phase. For identifying errors in the analytical phase, the approach proceeds in a second step after discarding regression outliers and it considers the residual standard error of the regression line as a measure of analytical variability. A comparison of this measure between laboratories helps identifying laboratories with high analytical variability. In a third and last step, regression lines with high analytical variability are discarded and the intercept and slope of the regression lines are considered as indicators of bias. They can be viewed as following a bivariate Normal distribution, and a robust approach can help identify regression lines with excessive bias. In comparison to other techniques that evaluate laboratory performance incorporating various samples [35, 93, 19, 188], this technique has demonstrated superior performance in estimating and handling outliers, and accurate estimation of bias and variability. This approach can also be applied for EQA surveys in which several samples, by preference more than six, have been sent out in one survey or for evaluation of laboratory performance over a long time period by combining results from different EQA surveys. In the former case, it yields estimates of intra-day accidental mistakes, variability and bias; in the latter case, it gives an overview of long-term laboratory performance.

The approach can also be adapted to specific aspects of various parameters. In case of lack of homogeneity of variances in the relation between target and reported values, a weighing factor can be applied for the linear regression model. In addition, when the distribution of the data around the target value is markedly skewed to be approximated by the Normal distribution, a logarithmic transformation of the data may be used before applying the method. The approach has been illustrated by EQA surveys for lithium and ethanol determination in blood, leucocyte subset counting, and semen analysis. For the three latter parameters, a long-term analysis was performed and similar conclusions were drawn for all surveys. Laboratory performance increased over time. The most obvious change was a decrease in accidental mistakes, identified as outliers against the linear regression model. A comparison of analytical variability and bias over time for laboratories that used the same analytical methodology over the study period, however, did not show any clear improvement, although performance differences between methods were observed. As a consequence, an improvement over time of laboratory performance for the parameters under consideration was explained by a decrease in accidental mistakes and an increased use of more accurate methods over time.

Reports that provide valuable feedback from EQA organizers to the participants can be considered as a useful tool to support the educational role of EQA [178, 60]. International standards emphasize the role of graphical representations in these reports. It is of increasing importance for EQA organizers to make deliberate decisions about the reporting of the data to their participants and use novel techniques to improve the communication channel with the laboratories. As far as the visual representation of data is concerned, various graphical techniques exist to display the reported data and to give the participants a visual appraisal of their position with respect to the other laboratories Traditional visual representations, such as Normal quantile plots [60, 87], histograms [60], box plots or empirical cumulative distribution curves can be applied, although in presence of outliers specific adaptations may be needed. Other visual representations that address specific questions may be given as well, such as plots that show the evolution of z-scores, or plots that combine bias and standard deviation in a single graph [48, 157, 60]. In addition, for the 3-step approach, a specific representation exists, using a scatter plot to show the linear regression line, an empirical distribution curve for displaying the regression residual standard error, and a scatter plot with a confidence ellipse for displaying the distribution of intercepts and slopes. Today, various technical solutions exist to display EQA data. For example, interactive graphs can be accessed via a web page. This allows the laboratories to find more textual information about a data point in a graph, or to quickly change some characteristic of the graphs (e.g. graphs including data obtained over several suveys, graphs where one can specify the time frame in which results should be seen). The graphs do not only allow EQA organizers to enrich their feedback to the laboratories, they also afford laboratories to explore their own results and help them to identify possible sources of mistakes.

In conclusion, while EQA organizers are facing new challenges and opportunities, the use of novel computing, statistical and graphical tools in EQA programmes can further enhance the quality and efficiency of clinical laboratory work.

This thesis is based on the following papers:

- Coucke, W., Devleeschouwer, N., Libeer, J.C., Schiettecatte, J., Martin, M. and Smitz, J. [2007], 'Accuracy and reproducibility of automated estradiol-17 and progesterone assays using native serum samples: results obtained in the Belgian external assessment scheme', Human Reproduction 22(12), 3204-3209.
- Van Blerk M., Coucke W., Chatelain B., Goossens W., Jochmans K., Meeus P., Mertens G., Pradier O., Rummens J.L., Scheiff J.M., Libeer J.C. [2007] External quality assessment in the measurement of haemoglobin by blood gas analysers in Belgium. Scandinavian Journal of Clinical & Laboratory Investigation 67(7), 735-740
- Van Blerk M., Van Campenhout C., Bossuyt X., Duchateau J., Humbel R., Servais G., Tomasi, J.P., Albert A., Coucke W., Libeer, J.C. [2008] Current practices in antinuclear antibody testing: results from the Belgian External Quality Assessment Scheme. Clinical Chemistry and Laboratory Medicine 47(1), 102-108
- Coucke, W., Van Blerk, M., Libeer, J.C., Van Campenhout, C. and Albert, A. [2010], 'A new statistical method for evaluating long-term analytical performance of laboratories applied to an external quality assessment scheme for flow cytometry', Clinical Chemistry and Laboratory Medicine 48(5), 645-650.

- Coucke W., China B., Delattre I., Lenga Y., Van Blerk M., Van Campenhout M., Van de Walle P., Vernelen K., Albert A. [2011] 'Comparison of different approaches to evaluate External Quality Assessment Data'. Clinica Chimica Acta 413(5–6), 582–586.
- Coucke W. and Soumali M.R. [2012] 'Interactive graphs for reporting results of External Quality Assessment schemes'. Scientific report 2010-2011 of the WIV-ISP, in press.

In addition four oral presentations were made:

- Coucke W. 'Accuraatheid en reproduceerbaarheid van serum estradiol- en progesteron metingen: een voorbeeld uit de Belgische Externe Kwaliteitsevaluatie' Presentation give at 'Postgraduaat en Navormingsprogramma Klinische Biologie 2007-2008', 29 November 2007, VUB Brussels
- Coucke W., Van Blerk M, Libeer JC, Albert A. 'A new statistical method for evaluating External Quality Assessment (EQA) data from flow cytometry' Presentation given at the 16th Annual Meeting of the Belgian Statistical Society, Namur, 16-17 October 2008
- Coucke W. 'A new statistical method for evaluating long terme analytical performance of laboratories, applied to an external quality assessment scheme for flow cytometry' Presentation given at 'Séminaire Ecole Doctorale Thématique Santé Publique, Santé et Société (EDT SPSS)', Journée Doctorale, Brussels, 18 November 2010.
- Coucke W, Albert A. 'Comparison of different z-score estimation methods for external quality assessment programs' Presentation given at the 19th Annual Meeting of the Belgian Statistical Society, Hasselt, 13-14 October 2011

Appendix

A1. Linear regression

Consider a response variable Y and an explanatory variable X. Let $\{(\mathrm{x_i},\mathrm{y_i}), \mathrm{i}=1,...\mathrm{n}\}$ be n observations of these variables. The linear regression model writes

$$\mathrm{E}\left(\mathrm{Y}|\mathrm{x}\right) = \mathrm{a} + \mathrm{b.x}$$

or

$$\mathrm{y_i} = \mathrm{a} + \mathrm{b.x_i} + \mathrm{e_i},\ \mathrm{i} = 1, .., \mathrm{n}$$

where a denotes the intercept, b the slope and $\mathrm{e_i}$ the residual or error term, assumed to be Normally distributed $\mathrm{N}(0,\sigma^2)$. Estimates of a and b are obtained by minimizing the residual sums of squares, namely

$$\mathrm{S} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e_i^2}$$

Note that minimizing S is the cornerstone of the regression line calculation. For this reason, linear regression is often called 'Least Squares Regression'.

Taking partial derivatives of the above equation with respect to a and b leads to the following results [152]

$$\begin{aligned}\hat{\mathrm{b}} &= \frac{\sum_{i=1}^{n}(y_i\text{-}\bar{y})(x_i\text{-}\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ \hat{\mathrm{a}} &= \bar{y}-b\bar{x}\end{aligned}$$

where $\bar{x}=\sum_{i=1}^{n}x_i/n$ and $\bar{y}=\sum_{i=1}^{n}y_i/n$.

Linear regression may also be calculated giving different weights to each pair (x_i, y_i). The formulas then become

$$\begin{aligned}\hat{\mathrm{b}} &= \frac{\sum_{i=1}^{n}w_i(y_i-\bar{y}_w)(x_i-\bar{x}_w)}{\sum_{i=1}^{n}w_i(x_i-\bar{x}_w)^2} \\ \hat{\mathrm{a}} &= \bar{y}-b.\bar{x}\end{aligned}$$

where $\bar{x}_w=\sum_{i=1}^{n}w_ix_i/\sum_{i=1}^{n}w_i$ and $\bar{y}_w=\sum_{i=1}^{n}w_iy_i/\sum_{i=1}^{n}w_i$. The values w_i are the weights for (x_i, y_i) and may be taken as the inverse of variance estimates.

The linear regression problem may also be written down in matrix notation. Disregarding weights, the relation between x and y is then given by

$$\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \ldots & \ldots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{bmatrix}$$

or, in shorter notation,

$$\utilde{Y} = \utilde{X}\utilde{B} + \utilde{E}$$

Solving for $\utilde{B}$ gives

$$\hat{\underset{\sim}{\mathrm{B}}} = (\underset{\sim}{\mathrm{X}}^T\underset{\sim}{\mathrm{X}})^{-1}\underset{\sim}{\mathrm{X}}^T\underset{\sim}{\mathrm{Y}}$$

The matrix $\underset{\sim}{\mathrm{X}}$ is also called the design matrix.

The estimator of σ^2 is given by

$$\mathrm{s}^2=\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\hat{\mathrm{e}}_{\mathrm{i}}^2}{\mathrm{n}-2}$$

where $\hat{\mathrm{e}}_{\mathrm{i}} = \mathrm{y}_{\mathrm{i}} - \left(\hat{\mathrm{a}} + \hat{\mathrm{b}}\mathrm{x}_{\mathrm{i}}\right)$.

Now, two analytical methods, represented by variables X and Y, are equal if the regression line is the 45°-line (a=0, b=1). Often, this test is assessed by building a hypothesis test for a and b separately

$$\mathrm{H}_0\text{: a=0} \leftrightarrow \mathrm{H}_1\text{: a}\neq 0$$

$$\mathrm{H}_0\text{: b=1} \leftrightarrow \mathrm{H}_1\text{: b}\neq 1$$

The solution is found by first calculating the standard errors of a and b:

$$\begin{aligned}\mathrm{SE}(\hat{\mathrm{a}}) &= \mathrm{s}\sqrt{\frac{1}{\mathrm{n}}+\frac{\bar{\mathrm{x}}^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_{\mathrm{i}}-\bar{\mathrm{x}})^2}}\\ \mathrm{SE}(\hat{\mathrm{b}}) &= \frac{\mathrm{s}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_{\mathrm{i}}-\bar{\mathrm{x}})^2}}\end{aligned}$$

Then, the null hypotheses are accepted if

$$\left|\frac{\hat{\mathrm{a}}}{\mathrm{SE}(\hat{\mathrm{a}})}\right| < \mathrm{Q}_t(1-\frac{\alpha}{2};n-2) \text{ and } \left|\frac{\hat{\mathrm{b}}-1}{\mathrm{SE}(\hat{\mathrm{b}})}\right| < \mathrm{Q}_t(1-\frac{\alpha}{2};n-2)$$

where $Q_t(1-\frac{\alpha}{2}; n-2)$ is the upper $\dfrac{\alpha}{2}$ quantile of the Student's t distribution with n-2 degrees of freedom and α is the probability of falsely rejecting the null hypothesis, by convention set at 0.05. Note that standard statistical software calculates the above equations by default. One should only be aware that the null hypothesis concerning the slope tested by default by many statistical software packages is whether b is equal to 0 and not to 1.

A2. Orthogonal regression

Assume that the available data ($\mathrm{y_i}$,$\mathrm{x_i}$) are observations of the true values ($\mathrm{Y_i}$,$\mathrm{X_i}$), i.e.

$$\mathrm{y_i} = \mathrm{Y_i} + \varepsilon_\mathrm{i}$$
$$\mathrm{x_i} = \mathrm{X_i} + \eta_\mathrm{i}$$

where $\mathrm{Y_i}$ and $\mathrm{X_i}$ (i=1,...,n) are distributed N(μ_y,α_y^2) and N(μ_x,α_x^2), respectively and ε_i and η_i are independent and distributed N(0,λ_ε^2) and N(0,λ_η^2), respectively. The linear model writes

$$\mathrm{E\,(Y|X) = a + b.X}$$

There is no simple algebraic solution to the problem of orthogonal regression, although some solutions may be given if assumptions are made. In clinical medicine one often assumes that the variabilities λ_ε^2 and λ_η^2 are known up to a fixed ratio δ, $\delta^2 = \frac{\lambda_\varepsilon^2}{\lambda_\eta^2}$. Then, the regression coefficients are calculated as

$$\begin{aligned} \hat{\mathrm{b}} &= \frac{\mathrm{s_{yy}} - \delta^2\mathrm{s_{xx}} + \sqrt{(\mathrm{s_{yy}} - \delta^2\mathrm{s_{xx}})^2 + 4\delta^2\mathrm{s_{xy}^2}}}{2.\mathrm{s_{xy}}} \\ \hat{\mathrm{a}} &= \bar{\mathrm{y}} - \hat{\mathrm{b}}.\bar{\mathrm{x}} \end{aligned}$$

where, respectively, $\mathrm{s_{xx}} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_i} - \bar{\mathrm{x}})^2$, $\mathrm{s_{xy}} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_i} - \bar{\mathrm{x}})(\mathrm{y_i} - \bar{\mathrm{y}})$ and $\mathrm{s_{yy}} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{y_i} - \bar{\mathrm{y}})^2$.

This approach is often called Deming regression. It assumes that the measurement uncertainty is constant over the whole measurement range. Since this assumption cannot always be made, one can also consider the weighted Deming regression [89]:

$$
\begin{aligned}
\hat{b} &= \frac{\sum_{i=1}^{n} w_i z_i y_i'}{\sum_{i=1}^{n} w_i z_i x_i'} \\
\\
\hat{a} &= \bar{y}_w - b.\bar{x}_w
\end{aligned}
$$

with $w_i = \frac{1}{\left[v_i + \hat{b}^2 u_i\right]}$, $u_i = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$,$v_i = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$,

$x_i' = x_i - \bar{x}_w$, $y_i' = y_i - \bar{y}_w$ and $z_i = w_i(\hat{v}_i x_i' + \hat{b}\hat{u}_i y_i')$.

Because z_i, w_i, $\bar{x}_w$ and $\bar{y}_w$ are functions of b, an iterative calculation procedure is required. First a value of n^{-1} can be given to every w_i. Next, estimates of z_i, b, $\bar{x}_w$ and $\bar{y}_w$ can be calculated, which can in turn be used to update the values of w_i. This iteration may converge until the new values of w_i obtained after each iteration don't differ more than a predefined small value. The sampling variability of the regression coefficients estimates, necessary to build hypothesis testing like in the simple linear regression case, are calculated as

$$
\begin{aligned}
\mathrm{Var}(\hat{b}) &= Q^2 \sum_{i=1}^{n} w_i^2 [(x_i')v_i + (y_i')u_i] \\
\mathrm{Var}(\hat{a}) &= \left(\sum_{i=1}^{n} w_i\right)^{-1} + 2(\bar{x}_w + 2\bar{z}_w Q) z_w Q + (\bar{x}_w + 2\bar{z}_w)\mathrm{Var}(\hat{b})
\end{aligned}
$$

where $Q^{-1} = \sum_{i=1}^{n} w_i[\frac{x_i' y_i'}{\hat{b}} + 4z_i'(z_i - x_i')]$ and $z_i' = z_i - \bar{z}_w$, $\bar{z}_w = \frac{\sum_{i=1}^{n} w_i z_i}{\sum_{i=1}^{n} w_i}$

Another assumption which can be made is the equality of the ratios of the analytical error variance to the total variance:

$$
\frac{\lambda_x^2}{\sigma_x^2} = \frac{\lambda_y^2}{\sigma_y^2}
$$

Then, the slope of the regression writes $\hat{\mathrm{b}} = \pm \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \bar{\mathrm{x}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \bar{\mathrm{y}})^2}$.
where the sign is equal to the correlation coefficient between X and Y. This approach is called the standardized principal component [38].

A3. Passing-Bablok regression

A3.1 Calculation of regression coefficients

A method that doesn't require any special assumptions regarding the distribution of the data or the measurement errors of the two methods was described by Passing and Bablok [9, 10].
Let

$$\begin{aligned} \mathrm{x_i} &= \mathrm{X_i} + \varepsilon_\mathrm{i} \\ \mathrm{y_i} &= \mathrm{Y_i} + \eta_\mathrm{i} \end{aligned}$$

where $\mathrm{X_i}$ and $\mathrm{Y_i}$ (i=1,...,n) are random variables from an arbitrary, continuous distribution and ε_i, η_i are random error terms, both coming from the same type of distribution. Their variances λ_ε^2 and λ_η^2 need not to be constant within the sampling range but should remain proportional, that is

$$\delta^2 = \frac{\lambda_\eta^2}{\lambda_\varepsilon^2}$$

The slopes of the straight lines between any two points are used for the estimation of b. They are given by $\mathrm{S_{ij}} = \frac{\mathrm{y_i - y_j}}{\mathrm{x_i - x_j}}$, $1 \leq \mathrm{i} < \mathrm{j} \leq \mathrm{n}$. There are $\mathrm{C_2^n}$ possible ways to connect any two points. Any $\mathrm{S_{ij}}$ with a value of -1, 0 or $\pm\infty$ is disregarded, so that in total there are $\mathrm{m} \leq \mathrm{C_2^n}$ slopes $\mathrm{S_{ij}}$. Observe that, since this method is only to be applied to continuous variables, the chance of having $\mathrm{S_{ij}}$ equal to -1, 0 or $\pm\infty$ is low.

Then, let k be the number of values of $\mathrm{S_{ij}}$ with $\mathrm{S_{ij}} < -1$ and $\mathrm{S_{(k)}}$ the kth slope in the sorted sequence of all the slopes, $\mathrm{S_{(1)}} \leq \mathrm{S_{(2)}} \leq \ldots \leq \mathrm{S_{(n)}}$. Using k as an offset, b is estimated by

$$\hat{b} = \begin{cases} S_{\left(\frac{m+1}{2} + k\right)} & \text{if m is odd} \\ \\ \frac{1}{2}\left(S_{\left(\frac{m}{2} + k\right)} + S_{\left(\frac{m}{2} + 1 + k\right)}\right) & \text{if m is even} \end{cases}$$

Remark that the formulas for the calculation of $\hat{b}$ differ slightly between the first [9] and last [10] paper of Passing and Bablok. We opted to display the formulas mentioned in the first paper.

A3.2 Confidence intervals for regression coefficients

For the construction of a two-sided confidence interval for b at the α level, let $Q_Z(1-\frac{\alpha}{2})$ denote the upper $\frac{\alpha}{2}$ quantile of the standardized Normal distribution.

With C_α=$Q_Z(1-\frac{\alpha}{2})\sqrt{\frac{n(n-1)(2n+5)}{18}}$ and $M_1 = \frac{m-C_\alpha}{2}$, $M_2 = m - M_1 + 1$, where M_1 is rounded to an integer value, the confidence interval for b is given by $S_{(M_1+k)} \leq b \leq S_{(M_2+k)}$.

The intercept a is given by

$$\hat{a} = \underset{1\leq i\leq n}{\text{med}} \left\{y_i - \hat{b}x_i\right\}$$

If b_L denotes the lower limit and b_U the upper limit of the confidence interval for b, then the corresponding limits for a are given by

$$\hat{a}_L = \underset{1\leq i\leq n}{\text{med}}\{y_i - \hat{b}_U x_i\}$$

$$\hat{a}_U = \underset{1\leq i\leq n}{\text{med}}\{y_i - \hat{b}_L x_i\}$$

It should be noted that the confidence intervals around the regression coefficients are usually wider than when classical linear regression is used [180, 9].

A4. Method comparison for multiple analyses of a limited number of samples

The assumption of linearity may be performed by a lack of fit test [105]. The test compares an ANOVA-model, where each group of measurements for a certain x_i as target value is considered as a separate group, with a regression model where the different values x_i are considered to be continuous. If there are n laboratories that have reported results for m samples, we can write the ANOVA and regression design matrices as

$$\text{ANOVA} \quad \begin{bmatrix} \mathrm{x_1} & 0 & 0 & \ldots & 0 \\ \ldots & 0 & 0 & \ldots & 0 \\ \mathrm{x_1} & 0 & 0 & \ldots & 0 \\ 0 & \mathrm{x_2} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \ldots & 0 \\ 0 & \mathrm{x_2} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \mathrm{x_m} \\ 0 & 0 & 0 & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \mathrm{x_m} \end{bmatrix}$$

$$\text{Regression} \quad \begin{bmatrix} 1 & \mathrm{x_1} \\ \ldots & \ldots \\ 1 & \mathrm{x_2} \\ \ldots & \ldots \\ 1 & \mathrm{x_m} \end{bmatrix}$$

In both matrices, there are as much values $\mathrm{x_i}$ as there are laboratories that have reported values for the corresponding sample. For both models, the residual sum of squares (SSE) is calculated.

For ANOVA, the Residual Sums of Squares writes $\mathrm{SSE_A}=\sum_{\mathrm{i=1}}^{\mathrm{m}}\sum_{\mathrm{j=1}}^{\mathrm{n}}(\mathrm{y_{ij}}-\bar{\mathrm{x}}_\mathrm{i})^2$ and for regression, we have $\mathrm{SSE_R}=\sum_{\mathrm{i=1}}^{\mathrm{m}}\sum_{\mathrm{j=1}}^{\mathrm{n}}(\mathrm{y_{ij}}-\hat{\mathrm{a}}-\hat{\mathrm{b}}\mathrm{x_i})^2$.

Then, the quantity

$$\mathrm{F} = \frac{SSE_R - SSE_A}{SSE_A} . \frac{n-m}{m-2}$$

follows an $\mathrm{F}_{(\mathrm{m}-2,\mathrm{n}-\mathrm{m})}$ distribution. The null hypothesis, stating that there is no lack of fit, is rejected if F is larger than a $Q_F(1-\alpha; m-2, n-m)$ where α is the false rejection rate. It is a very powerful test and may reject the null hypothesis solely based on statistical and no clinical reasons. For this reason, one may adopt it by allowing a minimal deviation from linearity for each sample and use the bootstrap [34] to obtain the null distribution.

A5. Multivariate Analysis of Variance

The matrix notation of a MANOVA model is similar to the notation of an ANOVA model:

$$\underset{\sim}{\mathrm{Y}}_{\mathrm{n\times p}} = \underset{\sim}{\mathrm{X}}_{\mathrm{n\times p}}.\underset{\sim}{\mathrm{B}}_{\mathrm{r\times p}} + \underset{\sim}{\mathrm{E}}_{\mathrm{r\times p}}$$

Assuming we have p samples and n laboratories, $\underset{\sim}{\mathrm{Y}}$ is an n×p matrix of response variables, $\underset{\sim}{\mathrm{X}}$ is a full-rank n×r matrix of group effects (for one-way MANOVA, r is the number of groups to compare), $\underset{\sim}{\mathrm{B}}$ is an r×p matrix of group effects. $\underset{\sim}{\mathrm{E}}$ is the n×p error matrix. The solution is found by calculating the between-group sum of squares and the within-group sum of squares matrix. The two matrices are used to calculate an F-statistic under the null hypothesis. Several versions of the test are available and there is no clear agreement about the optimality of the tests. We follow the recommendation of Olson [111, 110] and use the Pillai-Bartlett's trace, which is calculated as

$$\mathrm{PB} = \mathrm{tr}(\underset{\sim}{\mathrm{H}}\underset{\sim}{\mathrm{T}}^{-1})$$

where $\underset{\sim}{\mathrm{H}}$ is the between-group sum of squares matrix and reflects the differences between the groups. In general, its values increases with increasing difference between groups. $\underset{\sim}{\mathrm{T}}$ is the sum of the between-group and within-group sum of squares matrix, the latter reflecting the variability between values of the same group. The trace is the sum of the eigenvalues of a matrix. Given s=rank of $\underset{\sim}{\mathrm{H}}$, under the null hypothesis of no difference between groups, the Pillai-Bartlett's trace follows an F[p(r−1),s(n−r+s−p)] distribution. A hypothesis test compares PB with this F-distribution and rejects the null hypothesis when $\mathrm{PB}>\mathrm{Q_F}[1-\alpha;\mathrm{p(r-1)};\mathrm{s(n-r+s-p))}]$, where $\mathrm{s} = \mathrm{min(p, r-1)}$.

Bibliography

[1] Aakre, K., Thue, G., Subramaniam-Haavik, S., Bukve, T., Morris, H., Müller, M., Lovrencic, M., Plum, I., Kallion, K., Aab, A., Kutt, M., Gillery, P., Schneider, N., Horvath, A., Onody, R., Oosterhuis, W., Ricos, C., Perich, C., Nordin, G. and Sandverg, S. [2008], 'Postanalytical external quality assessment of urine albumin in primary health care: an international survey', *Clinical Chemistry* **54**(10), 1630–1636.

[2] Albert, A. [2011], Statistische methoden gebruikt voor EKE, Technical report.
URL: *https://www.wiv-isp.be/ClinBiol/bckb33/activities/external_quality/__down/_nl/Statistische_methoden_gebruikt_voor_EKE.pdf*

[3] Albert, A. and Harris, E. [1987], *Multivariate interpretation of clinical laboratory data*, Dekker (New York).

[4] Altman, D. G. and Bland, J. M. [1983], 'Measurement in medicine: the analysis of method comparison studies', *Journal of the Royal Statistical Society. Series D* **32**(3), 307–317.

[5] Anckaert, E., Mees, M., Schiettecatte, J. and Smitz, J. [2002], 'Clinical validation of a fully automated 17b-estradiol and progesterone assay

(VIDAS®) for use in monitoring assisted reproduction treatment', *Clinical Chemistry and Laboratory Medicine* **40**(8), 824–831.

[6] Armbruster, D. A. and Alexander, D. B. [2006], 'Sample to sample carryover: a source of analytical laboratory error and its relevance to integrated clinical chemistry/immunoassay systems', *Clinica Chimica Acta* **373**(1-2), 37–43.

[7] Atkinson, A. C. [1986], 'Masking unmasked', *Biometrika* **73**(3), 533 – 541.

[8] Baadenhuijsen, H., Steigstra, H., Cobbaert, C., Kuypers, A., Weykamp, C. and Jansen, R. [2002], 'Commutability assessment of potential reference materials using a multicenter split-patient-sample between-field-methods (twin-study) design: study within the framework of the dutch project "Calibration 2000"', *Clinical Chemistry* **48**(9), 1520–1525.

[9] Bablok, W. and Passing, H. [1983], 'A new biometrical procedure for testing the equality of measurements from two different analytical methods', *J. Clin. Chem. Clin. Biochem* **21**, 709–720.

[10] Bablok, W., Passing, H., Bender, R. and Schneider, B. [1988], 'A general regression procedure for method transformation. application of linear regression procedures for method comparison studies in clinical chemistry, part III', *Clinical Chemistry and Laboratory Medicine* **26**(11), 783–790.

[11] Bieglmayer, C., Chan, D. W., Sokoll, L., Imdahl, R., Kobayashi, M., Yamada, E., Lilje, D. J., Luthe, H., Meissner, J., Messeri, G., Celli, A., Tozzi, P., Roth, H., Schmidt, F., Mächler, M., Schuff-Werner, P., Zingler, C., Smitz, J., Schiettecatte, J., Vonderschmitt, D. J., Pei, P., Ng, K., Ebert, C., Kirch, P., Wanger, M., McGovern, M., Stockmann, W. and Kunst, A. [2004], 'Multicentre performance evaluation of the e170 module for modular analytics', *Clinical Chemistry and Laboratory Medicine* **42**(10), 1186–1202.

[12] Bland, J. M. and Altman, D. G. [1999], 'Measuring agreement in method comparison studies', *Statistical methods in medical research* **8**(2), 135.

[13] Bonini, P., Ceriotti, F., Mirandola, G. and Signori, C. [2008], 'Misidentification and other preanalytical errors', *Journal of Medical Biochemistry* **27**(3), 339–342.

[14] Bonini, P., Plebani, M., Ceriotti, F. and Rubboli, F. [2002], 'Errors in laboratory medicine', *Clinical Chemistry* **48**(5), 691–698.

[15] Boone, D. J. [2004], 'Is it safe to have a laboratory test?', *Accreditation and Quality Assurance* **10**(1), 5–9.

[16] Bossuyt, X., Verweire, K. and Blanckaert, N. [2007], 'Laboratory medicine: challenges and opportunities', *Clinical chemistry* **53**(10), 1730.

[17] Boudou, P., Taieb, J., Mathian, B., Badonnel, Y., Lacroix, I., Mathieu, E., Millot, F., Queyrel, N., Somma-Delpero, C. and Patricot, M. C. [2001], 'Comparison of progesterone concentration determination by 12 non-isotopic immunoassays and gas chromatography/mass spectrometry in 99 human serum samples', *The Journal of Steroid Biochemistry and Molecular Biology* **78**(1), 97–104.

[18] Bremser, W., Lücke, F., Urmetzer, C., Fuchs, E. and Leist, U. [2011], 'An approach to integrated data assessment in a proficiency test on the enumeration of escherichia coli', *Journal of Applied Microbiology* **110**(1), 128–138.

[19] Bullock, D. and Wilde, C. [1985], 'External quality assessment of urinary pregnancy oestrogen assay: further experience in the United Kingdom', **22**, 273–82.

[20] Burke, M. D. [2000], 'Laboratory medicine in the 21st century', *American Journal of Clinical Pathology* **114**(6), 841–846.

[21] Carraro, P. and Plebani, M. [2007], 'Errors in a stat laboratory: types and frequencies 10 years later', *Clinical chemistry* **53**(7), 1338.

[22] Cattozzo, G., Albeni, C., Franzini, C. et al. [2010], 'Harmonization of values for serum alkaline phosphatase catalytic activity concentration employing commutable calibration materials', *Clinica Chimica Acta* **411**(11-12), 882–885.

[23] Check, J. H., Ubelacker, L. and Lauer, C. C. [1995], 'Falsely elevated steroidal assay levels related to heterophile antibodies against various animal species', *Gynecologic and Obstetric Investigation* **40**(2), 139–140.

[24] Clinical and Laboratory Standards Institute [1999], 'Preparation and validation of commutable frozen human serum pools as secondary reference materials for cholesterol measurement procedures; approved guideline. CLSI document C-37A [1999]'.

[25] Cooper, G. R., Myers, G. L., Kimberly, M. M. and Waymack, P. P. [2002], 'The effects of errors in lipid measurement and assessment', *Current Cardiology Reports* **4**(6), 501–507.

[26] Coucke, W., China, B., Delattre, I., Lenga, Y., Van Blerk, M., Van Campenhout, C., Van de Walle, P., Vernelen, K. and Albert, A. [2012], 'Comparison of different approaches to evaluate External Quality Assessment Data', *Clinica Chimica Acta* **413**(5–6), 582–586.

[27] Coucke, W., Devleeschouwer, N., Libeer, J., Schiettecatte, J., Martin, M. and Smitz, J. [2007], 'Accuracy and reproducibility of automated estradiol-17β and progesterone assays using native serum samples: results obtained in the belgian external assessment scheme', *Human Reproduction* **22**(12), 3204 –3209.

[28] Coucke, W., Van Blerk, M., Libeer, J., Campenhout, C. V. and Albert, A. [2010], 'A new statistical method for evaluating long-term analytical performance of laboratories applied to an external quality assessment scheme for flow cytometry', *Clinical Chemistry and Laboratory Medicine* **48**(5), 645–650.

[29] Davey, D. D., McGoogan, E., Somrak, T. M., Allen, K. A., Beccati, D., Cramer, S. F., Frable, W. J., Hauser, N. J., Hewer, E. M., Lestadi,

J., Lulla, M. K., O'Rourke, D. and Suprun, H. Z. [2000], 'Competency assessment and proficiency testing', *Acta Cytologica* **44**(6), 939–943.

[30] Dixon, W. J. [1950], 'Analysis of extreme values', *The Annals of Mathematical Statistics* **21**(4), 488–506.

[31] Dixon, W. J. [1951], 'Ratios involving extreme values', *The Annals of Mathematical Statistics* **22**(1), 68–78.

[32] Duewer, D. [2008], 'A comparison of location estimators for interlaboratory data contaminated with value and uncertainty outliers', *Accreditation and Quality Assurance* **13**(4-5), 193–216.

[33] Eckfeldt, J. H. and Copeland, K. R. [1993], 'Accuracy verification and identification of matrix effects. the College of American Pathologists' Protocol', *Archives of Pathology & Laboratory Medicine* **117**(4), 381–386.

[34] Efron, B. and Tibshirani, R. [1994], *An Introduction to the Bootstrap*, 1 edn, Chapman and Hall/CRC.

[35] Ehrmeyer, S. and Laessig, R. [1985], 'Alternative statistical approach to evaluating interlaboratory performance', *Clin Chem* **31**(1), 106–108.

[36] European Parliament and Council [1998], 'Directive 98/79/EC of the european parliament and of the council of 27 october 1998 on in vitro diagnostic medical devices', *Official Journal of the European Union L* **41**, 1–37.

[37] Felder, R. A. [2011], 'Preanalytical errors introduced by sample-transportation systems: a means to assess them', *Clinical Chemistry* **57**(10), 1349–1350.

[38] Feldmann, U., Schneider, B., Klinkers, H. and Haeckel, R. [1981], 'A multivariate approach for the biometric comparison of analytical methods in clinical chemistry', *Clinical Chemistry and Laboratory Medicine* **19**(3), 121–138.

[39] Férard, G., Edwards, J., Kanno, T., Lessinger, J. M., Moss, D. W., Schiele, F., Tietz, N. W. and Vassault, A. [1998], 'Interassay calibration as a major contribution to the comparability of results in clinical enzymology', *Clinical Biochemistry* **31**(6), 489–494.

[40] Ferrero, C. A., Carobene, A., Ceriotti, F., Modenese, A. and Arcelloni, C. [1995], 'Behavior of frozen serum pools and lyophilized sera in an external quality-assessment scheme', *Clinical Chemistry* **41**(4), 575–580.

[41] Franzini, C. and Ceriotti, F. [1998], 'Impact of reference materials on accuracy in clinical chemistry', *Clinical Biochemistry* **31**(6), 449–457.

[42] Friedecky, B., Kratochvila, J. and Budina, M. [2011], 'Why do different EQA schemes have apparently different limits of acceptability?', *Clinical Chemistry and Laboratory Medicine* **49**(4), 743–745.

[43] Goldie, D. J. [2001], 'Accreditation of external quality assessment schemes in the united kingdom', *Clinica Chimica Acta* **309**(2), 179–181.

[44] Grubbs, F. E. [1969], 'Procedures for detecting outlying observations in samples', *Technometrics* **11**(1), 1–21.

[45] Guder, W. G. and Buttner, J. [1997], 'Clinical chemistry in laboratory medicine in Europe-Past, present and future challenges', *European journal of clinical chemistry and clinical biochemistry* **35**, 487–494.

[46] Guidi, G. C. and Lippi, G. [2006], 'Laboratory medicine in the 2000s: programmed death or rebirth?', *Clinical Chemistry and Laboratory Medicine* **44**(8), 913–917.

[47] Haeckel, R. and Wosniok, W. [2011], 'A new concept to derive permissible limits for analytical imprecision and bias considering diagnostic requirements and technical state-of-the-art', *Clinical Chemistry and Laboratory Medicine* **49**(4), 623–635.

[48] Harms, A. V. [2009], 'Visualisation of proficiency test exercise results in Kiri plots', *Accreditation and Quality Assurance* **14**(6), 307–311.

[49] Harris, E. K. and Boyd, J. C. [1995], *Statistical bases of reference values in laboratory medicine*, Vol. 146, Marcel Dekker, New York; NY, US.

[50] Healy, M. [1979], 'Outliers in clinical chemistry quality-control schemes', *Clin Chem* **25**(5), 675–677.

[51] Hendriks, H., Kortlandt, W. and Verweij, W. [2000], 'Analytical performance comparison of five new generation immunoassay analyzers', *Ned Tijdschr Klin Chem* **25**(3), 170–177.

[52] Henriksen, G. M., Pedersen, M. M., Norgaard, I., Blom, M., Blou, L., Blaabjerg, O. and Uldall, A. [2004], 'Minimally processed fresh frozen human reference sera: preparation, testing, and application to international external quality assurance', *Scandinavian Journal of Clinical & Laboratory Investigation* **64**(4), 293–308.

[53] Hertzberg, M. S., Mammen, J., McCraw, A., Nair, S. C. and Srivastava, A. [2006], 'Achieving and maintaining quality in the laboratory', *Haemophilia: The Official Journal of the World Federation of Hemophilia* **12 Suppl 3**, 61–67.

[54] Hilborne, L. H., Lubin, I. M. and Scheuner, M. T. [2009], 'The beginning of the second decade of the era of patient safety: Implications and roles for the clinical laboratory and laboratory professionals', *Clinica Chimica Acta* **404**(1), 24–27.

[55] Hollensead, S. C., Lockwood, W. B. and Elin, R. J. [2004], 'Errors in pathology and laboratory medicine: consequences and prevention', *Journal of surgical oncology* **88**(3), 161–181.

[56] Howanitz, P. J. and Cembrowski, G. S. [2000], 'Postanalytical quality improvement: a college of american pathologists Q-Probes study of elevated calcium results in 525 institutions', *Archives of pathology & laboratory medicine* **124**(4), 504–510.

[57] Hund, E., Massart, D. L. and Smeyers-Verbeke, J. [2000], 'Interlaboratory studies in analytical chemistry', *Analytica Chimica Acta* **423**(2), 145–165.

[58] International Organization for Standardization [2003*a*], 'ISO 17511:2003. In vitro diagnostic medical devices - measurement of quantities in biological samples - metrological traceability of values assigned to calibrators and control materials'.

[59] International Organization for Standardization [2003*b*], 'ISO 18153:2003. In vitro diagnostic medical devices - measurement of quantities in biological samples - metrological traceability of values for catalytic concentration of enzymes assigned calibrators and control materials'.

[60] International Organization for Standardization [2005*a*], 'ISO 13528:2005. Statistical methods for use in proficiency testing by interlaboratory comparisons'.

[61] International Organization for Standardization [2005*b*], 'ISO/IEC 17025: 2005. General requirements for the competence of testing and calibation laboratories'.

[62] International Organization for Standardization [2007], 'ISO 15189: 2007. medical laboratories - particular requirements for quality and competence'.

[63] Isenberg, H. D. and D'Amato, R. F. [1996], 'Does proficiency testing meet its objective?', *Journal of Clinical Microbiology* **34**(11), 2643–2644.

[64] Jain, R. B. [2010], 'A recursive version of Grubbs' test for detecting multiple outliers in environmental and chemical data', *Clinical Biochemistry* **43**(12), 1030–1033.

[65] Jay, D. W. [2011], 'Method comparison: where do we draw the line?', *Clinical Chemistry and Laboratory Medicine* **49**(7), 1089–1090.

[66] Jurasović, J., Cvitković, P., Pizent, A., Čolak, B. and Telišman, S. [2004], 'Semen quality and reproductive endocrine function with regard to blood cadmium in croatian male subjects', *BioMetals* **17**(6), 735–743.

[67] Kallner, A., McQueen, M. and Heuck, C. [1999], 'The Stockholm Consensus Conference on quality specifications in laboratory medicine, 25-26 april 1999', *Scandinavian Journal of Clinical and Laboratory Investigation* **59**(7), 475–476.

[68] Kalra, J. [2004], 'Medical errors: impact on clinical laboratories and other critical areas', *Clinical biochemistry* **37**(12), 1052–1062.

[69] Kettelhut, M. M., Chiodini, P. L., Edwards, H. and Moody, A. [2003], 'External quality assessment schemes raise standards: evidence from the UKNEQAS parasitology subschemes', *Journal of Clinical Pathology* **56**(12), 927–932.

[70] Kilpatrick, E. [2004], 'Can the addition of interpretative comments to laboratory reports influence outcome? an example involving patients taking thyroxine', *Ann Clin Biochem* **41**(3), 227–229.

[71] Kohn, L. T., Corrigan, J. M., Donaldson, M. S. et al. [1999], 'To err is human: building a safer health system', *Washington, DC* .

[72] Kricka, L. J. [1999], 'Human Anti-Animal antibody interferences in immunological assays', *Clin Chem* **45**(7), 942–956.

[73] Kristoffersen, A., Thue, G. and Sandberg, S. [2006], 'Postanalytical external quality assessment of warfarin monitoring in primary healthcare', *Clinical Chemistry* **52**(10), 1871–1878.

[74] Krouwer, J. S. and Cembrowski, G. S. [2011], 'Towards more complete specifications for acceptable analytical performance - a plea for error grid analysis', *Clinical Chemistry and Laboratory Medicine* **49**(7), 1127–1130.

[75] Lasky, F. D. [1993], 'Achieving accuracy for routine clinical chemistry methods by using patient specimen correlations to assign calibrator val-

ues. a means of managing matrix effects', *Archives of Pathology & Laboratory Medicine* **117**(4), 412–419.

[76] Levey, S. and Jennings, E. R. [1950], 'The use of control charts in the clinical laboratory', *American Journal of Clinical Pathology* **20**(11), 1059–1066.

[77] Libeer, J. C. [2004], 'Wood WG. questionable results–who directs the EQAS organisers?', *Clinical Chemistry and Laboratory Medicine* **42**(9), 1074–1075.

[78] Lim, E. M., Sikaris, K. A., Gill, J., Calleja, J., Hickman, P. E., Beilby, J. and Vasikaran, S. D. [2004], 'Quality assessment of interpretative commenting in clinical chemistry', *Clin Chem* **50**(3), 632–637.

[79] Lippi, G. [2009], 'Governance of preanalytical variability: Travelling the right path to the bright side of the moon?', *Clinica Chimica Acta* **404**(1), 32–36.

[80] Lippi, G., Bassi, A., Brocco, G., Montagnana, M., Salvagno, G. L. and Guidi, G. C. [2006], 'Preanalytic error tracking in a laboratory medicine department: results of a 1-year experience', *Clinical chemistry* **52**(7), 1442.

[81] Lippi, G., Blanckaert, N., Bonini, P., Green, S., Kitchen, S., Palicka, V., Vassault, A. J., Mattiuzzi, C. and Plebani, M. [2009], 'Causes, consequences, detection, and prevention of identification errors in laboratory diagnostics', *Clinical Chemistry and Laboratory Medicine* **47**(2), 143–153.

[82] Lippi, G., Chance, J. J., Church, S., Dazzi, P., Fontana, R., Giavarina, D., Grankvist, K., Huisman, W., Kouri, T., Palicka, V., Plebani, M., Puro, V., Salvagno, G. L., Sandberg, S., Sikaris, K., Watson, I., Stankovic, A. K. and Simundic, A. [2011], 'Preanalytical quality improvement: from dream to reality', *Clinical Chemistry and Laboratory Medicine* **49**(7), 1113–1126.

[83] Lippi, G., Guidi, G. C., Mattiuzzi, C. and Plebani, M. [2006], 'Preanalytical variability: the dark side of the moon in laboratory testing', *Clinical Chemistry and Laboratory Medicine* **44**(4), 358–365.

[84] Lippi, G., Siest, G. and Plebani, M. [2008], 'Pharmacy-based laboratory services: past or future and risk or opportunity?', *Clinical Chemistry and Laboratory Medicine* **46**(4), 435–436.

[85] Loh, T. P., Saw, S., Chai, V. and Sethi, S. K. [2011], 'Impact of phlebotomy decision support application on sample collection errors and laboratory efficiency', *Clinica Chimica Acta* **412**(3-4), 393–395.

[86] Long, T. [1993], 'Statistical power in the detection of matrix effects', *Archives of Pathology & Laboratory Medicine* **117**(4), 387–392.

[87] Lowthian, P. J. and Thompson, M. [2002], 'Bump-hunting for the proficiency tester-searching for multimodality', *Analyst* **127**(10), 1359–1364.

[88] Lundberg, G. D. [1981], 'Acting on significant laboratory results', *JAMA: The Journal of the American Medical Association* **245**(17), 1762–1763.

[89] Martin, R. F. [2000], 'General deming regression for estimating systematic bias and its confidence interval in Method-Comparison studies', *Clin Chem* **46**(1), 100–104.

[90] Massart, C., Gibassier, J., Laurent, M. and Lannou, D. L. [2006], 'Analytical performance of a new two-step (Advia®) centaur estradiol immunoassay during ovarian stimulation', *Clinical Chemistry and Laboratory Medicine* **44**(1), 105–109.

[91] McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook, J., Williams, W., Case, J. and Maloney, P. [2003], 'LOINC, a universal standard for identifying laboratory observations: A 5-Year update', *Clin Chem* **49**(4), 624–633.

[92] Meijer, P. [2004], 'Wood WG. questionable results–who directs the EQAS organisers? a comment from the point of view of the ECAT foundation', *Clinical Chemistry and Laboratory Medicine* **42**(9), 1081–1081.

[93] Meijer, P., de Maat, M., Kluft, C., Haverkate, F. and van Houwelingen, H. [2002], 'Long-Term analytical performance of hemostasis field methods as assessed by evaluation of the results of an external quality assessment program for antithrombin', *Clin Chem* **48**(7), 1011–1015.

[94] Meijer, P. and Haverkate, F. [2005], 'External quality assessment and the laboratory diagnosis of thrombophilia', *Seminars in Thrombosis and Hemostasis* **31**(01), 59–65.

[95] Meijer, P. and Haverkate, F. [2006], 'An external quality assessment program for von willebrand factor laboratory analysis: an overview from the European concerted action on thrombosis and disabilities foundation', *Seminars in Thrombosis and Hemostasis* **32**(5), 485–491.

[96] Meijer, P., Kluft, C., Haverkate, F. and De Maat, M. P. M. [2003], 'The long-term within- and between-laboratory variability for assay of antithrombin, and proteins c and s: results derived from the external quality assessment program for thrombophilia screening of the ECAT foundation', *Journal of Thrombosis and Haemostasis* **1**(4), 748–753.

[97] Menditto, A., Patriarca, M. and Magnusson, B. [2006], 'Understanding the meaning of accuracy, trueness and precision', *Accreditation and Quality Assurance* **12**(1), 45–47.

[98] Miller, W. G. [2003], 'Specimen materials, target values and commutability for external quality assessment (proficiency testing) schemes', *Clinica Chimica Acta* **327**(1-2), 25–37.

[99] Miller, W. G. [2006], 'Why commutability matters', *Clinical Chemistry* **52**(4), 553–554.

[100] Miller, W. G. [2009], 'The role of proficiency testing in achieving standardization and harmonization between laboratories', *Clinical biochemistry* **42**(4-5), 232–235.

[101] Miller, W. G., Erek, A., Cunningham, T. D., Oladipo, O., Scott, M. G. and Johnson, R. E. [2011], 'Commutability limitations influence quality control results with different reagent lots', *Clinical Chemistry* **57**(1), 76–83.

[102] Miller, W. G., Myers, G. L., Ashwood, E. R., Killeen, A. A., Wang, E., Ehlers, G. W., Hassemer, D., Lo, S. F., Seccombe, D., Siekmann, L., Thienpont, L. M. and Toth, A. [2008], 'State of the art in trueness and interlaboratory harmonization for 10 analytes in general clinical chemistry', *Archives of Pathology & Laboratory Medicine* **132**(5), 838–846.

[103] Miller, W. G., Myers, G. L., Lou Gantzer, M., Kahn, S. E., Schönbrunner, E. R., Thienpont, L. M., Bunk, D. M., Christenson, R. H., Eckfeldt, J. H., Lo, S. F., Nübling, C. M. and Sturgeon, C. M. [2011], 'Roadmap for harmonization of clinical laboratory measurement procedures', *Clinical Chemistry* **57**(8), 1108–1117.

[104] Muller, K. E. and Peterson, B. L. [1984], 'Practical methods for computing power in testing the multivariate general linear hypothesis', *Computational Statistics & Data Analysis* **2**(2), 143–158.

[105] Neter, J., Kutner, M., Wasserman, W., Nachtsheim, C. and Neter, J. [1996], *Applied Linear Statistical Models*, 4 edn, McGraw-Hill/Irwin.

[106] Nevalainen, D., Berte, L., Kraft, C., Leigh, E., Picaso, L. and Morgan, T. [2000], 'Evaluating laboratory performance on quality indicators with the six sigma scale', *Archives of Pathology & Laboratory Medicine* **124**(4), 516–519.

[107] Noble, M. A. [2002], 'Advances in microbiology EQA', *Accreditation and Quality Assurance* **7**(8), 341–344.

[108] Noble, M. A. [2007], ‘Does external evaluation of laboratories improve patient safety?’, *Clinical Chemistry and Laboratory Medicine* **45**(6), 753–755.

[109] O’Brien, R. G., Muller, K. E. and Keith, E. [1993], Unified power analysis for t-tests through multivariate hypotheses., *in* ‘Applied analysis of variance in behavioral science’, Vol. 137 of *Statistics: Textbooks and monographs*, Marcel Dekker, New York; NY, US, pp. 297–344.

[110] Olson, C. [1976], ‘On choosing a test statistic in multivariate analysis of variance’, *Psychological Bulletin* **83**(4), 579–586.

[111] Olson, C. [1979], ‘Practical considerations in choosing a MANOVA test statistic: A rejoinder to stevens’, *Psychological Bulletin* **86**(6), 1350–1352.

[112] Panteghini, M. [2004], ‘The future of laboratory medicine: understanding the new pressures’, *The Clinical Biochemist Reviews* **25**(4), 207.

[113] Panteghini, M. [2007], ‘Traceability, reference systems and result comparability’, *The Clinical Biochemist Reviews* **28**(3), 97–104.

[114] Panteghini, M. [2009], ‘Traceability as a unique tool to improve standardization in laboratory medicine’, *Clinical biochemistry* **42**(4-5), 236–240.

[115] Panteghini, M. and Forest, J. [2005], ‘Standardization in laboratory medicine: New challenges’, *Clinica Chimica Acta* **355**(1-2), 1–12.

[116] Petersen, P. and Fraser, C. [2010], ‘Strategies to set global analytical quality specifications in laboratory medicine: 10 years on from the stockholm consensus conference’, *Accreditation and Quality Assurance* **15**(6), 323–330.

[117] Petersen, P. H., Fraser, C., Kallner, A. and (eds) Kenny, D. [1999], ‘Strategies to set global analytical quality specifications in laboratory medicine’, *Scandinavian Journal of Clinical and Laboratory Investigation* **59**, 475–585.

[118] Pizent, A., Čolak, B., Kljaković, Z. and Telišman, S. [2009], 'Prostate-Specific antigen (PSA) in serum in relation to blood lead concentration and alcohol consumption in men', *Archives of Industrial Hygiene and Toxicology* **60**(1), 69–78.

[119] Plebani, M. [1999], 'The clinical importance of laboratory reasoning', *Clinica Chimica Acta* **280**(1-2), 35–45.

[120] Plebani, M. [2004], 'Pre and post examination aspects', *Journal of the International International Federation of Clinical Chemistry* **15**(4).

[121] Plebani, M. [2005], 'External quality assessment programs: Past, present and future', *Jugoslovenska medicinska biohemija* **24**(3), 201–206.

[122] Plebani, M. [2006], 'Errors in clinical laboratories or errors in laboratory medicine?', *Clinical Chemistry and Laboratory Medicine* **44**(6), 750–759.

[123] Plebani, M. [2007*a*], 'Errors in laboratory medicine and patient safety: the road ahead', *Clinical Chemistry and Laboratory Medicine* **45**(6), 700–707.

[124] Plebani, M. [2007*b*], 'Quality specifications: self pleasure for clinical laboratories or added value for patient management?', *Clinical Chemistry and Laboratory Medicine* **45**(4), 462–466.

[125] Plebani, M. [2009], 'Exploring the iceberg of errors in laboratory medicine', *Clinica Chimica Acta* **404**(1), 16–23.

[126] Plebani, M. and Carraro, P. [1997], 'Mistakes in a stat laboratory: types and frequency', *Clinical Chemistry* **43**(8 Pt 1), 1348–1351.

[127] Plebani, M. and Lippi, G. [2011], 'Closing the brain-to-brain loop in laboratory testing', *Clinical Chemistry and Laboratory* **49**(7), 1131–1133.

[128] Plebani, M. and Piva, E. [2010], 'Medical errors: Pre-analytical issue in patient safety', *Journal of Medical Biochemistry* **29**(4), 310–314.

[129] Plum, I., Jörgensen, N. and Möller, J. [2004], 'Wood WG. questionable results–who directs the EQAS organisers?', *Clinical Chemistry and Laboratory Medicine* **42**(9), 1076–1077.

[130] Rej, R. [1994], 'Proficiency testing, matrix effects, and method evaluation', *Clinical Chemistry* **40**(3), 345–346.

[131] Rej, R. [2002], 'Proficiency testing and external quality assurance: crossing borders and disciplines', *Accreditation and Quality Assurance* **7**(8-9), 335–340.

[132] Rej, R. and Jenny, R. W. [1992], 'How good are clinical laboratories? an assessment of current performance', *Clinical Chemistry* **38**(7), 1210–1217; discussion 1218–1225.

[133] Ricós, C., García-Victoria, M. and de la Fuente, B. [2004], 'Quality indicators and specifications for the extra-analytical phases in clinical laboratory management', *Clinical chemistry and laboratory medicine* **42**(6), 578–582.

[134] Ricós, C., Juvany, R., Alvarez, V., Jiménez, C. V., Perich, C., Minchinela, J., Hernández, A. and Simón, M. [1997], 'Commutability between stabilized materials and fresh human serum to improve laboratory performance', *Clinica Chimica Acta* **263**(2), 225–238.

[135] Ricós, C., Juvany, R., Jiménez, C. V., Perich, C., Minchinela, J., Hernández, A., Simón, M. and Alvarez, V. [1997], 'Procedure for studying commutability validated by biological variation', *Clinica Chimica Acta* **268**(1-2), 73–83.

[136] Ricós, C., Juvany, R., Simón, M., Hernández, A., Alvarez, V., Jiménez, C. V., Minchinela, J. and Perich, C. [1999], 'Commutability and traceability: their repercussions on analytical bias and inaccuracy', *Clinica Chimica Acta* **280**(1-2), 135–145.

[137] Rin, G. D. [2010], 'Pre-analytical workstations as a tool for reducing laboratory errors', *Journal of Medical Biochemistry* **29**(4), 315–324.

[138] Rodríguez-Espinosa, J., Otal-Entraigas, C., Gascón-Roche, N., Mora-Brugués, J., Urgell-Rull, E., Bordás-Serrat, J. and Viscasillas-Molins, P. [1998], 'Analytical and clinical performance of an automated immunoassay system (Immulite®) for estradiol in serum', *Clinical Chemistry and Laboratory Medicine* **36**(12), 969–974.

[139] Rosner, B. [1975], 'On the detection of many outliers', *Technometrics* **17**(2), 221–227.

[140] Rousseeuw, P. J. and Croux, C. [1993], 'Alternatives to the median absolute deviation', *Journal of the American Statistical Association* **88**(424), 1273–1283.

[141] Rousseeuw, P. and Leroy, A. [1987], *Robust regression and outlier detection*, John Wiley and Sons.

[142] Rousseeuw, P. and van Driessen, K. [1999], 'A fast algorithm for the minimum covariance determinant estimator', *Technometrics* **41**(3), 212–223.

[143] Rousseeuw, P. and Zomeren, B. [1990], 'Unmasking multivariate outliers and leverage points', *Journal of the American Statistical Association* **85**(411), 633–639.

[144] Salfinger, M. and Ahmedov, S. [2009], 'Has the time come to discontinue proficiency testing?', *The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease* **13**(10), 1193.

[145] Salinas, M., López-Garrigós, M. and Uris, J. [2011], 'Towards laboratory knowledge, not data, in 70% of clinical decision-making. what "knowledge management" can add to clinical practice?', *Clinical Chemistry and Laboratory Medicine* **49**(8), 1389–1390.

[146] Schleicher, E. [2006], 'The clinical chemistry laboratory: current status, problems and diagnostic prospects', *Analytical and bioanalytical chemistry* **384**(1), 124–131.

[147] Sciacovelli, L., O'Kane, M., Skaik, Y. A., Caciagli, P., Pellegrini, C., Da Rin, G., Ivanov, A., Ghys, T. and Plebani, M. [2011], 'Quality indicators in laboratory medicine: from theory to practice. preliminary data from the IFCC working group project "Laboratory errors and patient safety"', *Clinical Chemistry and Laboratory Medicine* **49**(5), 835–844.

[148] Sciacovelli, L., Secchiero, S., Zardo, L. and Plebani, M. [2001], 'External quality assessment schemes: need for recognised requirements', *Clinica Chimica Acta* **309**(2), 183–199.

[149] Sciacovelli, L., Secchiero, S., Zardo, L., Zaninotto, M. and Plebani, M. [2006], 'External quality assessment: an effective tool for clinical governance in laboratory medicine', *Clinical Chemistry and Laboratory Medicine* **44**(6), 740–749.

[150] Sciacovelli, L., Zardo, L., Secchiero, S. and Plebani, M. [2004], 'Quality specifications in EQA schemes: from theory to practice', *Clinica Chimica Acta* **346**(1), 87–97.

[151] Secchiero, S., Sciacovelli, L., Zardo, L. and Plebani, M. [2005], 'Reply to WG wood. questionable results–who directs the EQAS organisers? clin chem lab med 2004; 42: 1073', *Clinical Chemistry and Laboratory Medicine* **43**(3), 346–348.

[152] Sen, A. K. and Srivastava, M. S. [1997], *Regression analysis: theory, methods and applications*, Springer.

[153] Siftar, Z., Paro, M. M. K., Sokolić, I., Nazor, A. and Mestrić, Z. F. [2010], 'External quality assessment in clinical cell analysis by flow cytometry. Why is it so important?', *Collegium Antropologicum* **34**(1), 207–217.

[154] Siloaho, M., Linko, S., Puhakainen, E. and Nordberg, U. R. [2006], 'The effects of quality-management systems on external quality-assessment performance in Finnish clinical chemistry laboratories', *Accreditation and Quality Assurance* **11**(5), 238–245.

[155] Skeie, S., Perich, C., Ricos, C., Araczki, A., Horvath, A. R., Oosterhuis, W. P., Bubner, T., Nordin, G., Delport, R., Thue, G. and Sandberg, S. [2005], 'Postanalytical external quality assessment of blood glucose and hemoglobin A1c: an international survey', *Clinical Chemistry* **51**(7), 1145–1153.

[156] Sonntag, O. [2009], 'Analytical interferences and analytical quality', *Clinica Chimica Acta* **404**(1), 37–40.

[157] Spasova, Y., Pommé, S. and Wätjen, U. [2007], 'Visualisation of interlaboratory comparison results in PomPlots', *Accreditation and Quality Assurance* **12**(12), 623–627.

[158] Steurer, J., Fischer, J. E., Bachmann, L. M., Koller, M. and Riet, G. T. [2002], 'Communicating accuracy of tests to general practitioners: a controlled study', *Bmj* **324**(7341), 824.

[159] Stöckl, D., Dewitte, K. and Thienpont, L. M. [1998], 'Validity of linear regression in method comparison studies: is it limited by the statistical model or the quality of the analytical input data?', *Clinical Chemistry* **44**(11), 2340–2346.

[160] Stöckl, D., Libeer, J. C., Reinauer, H., Thienpont, L. M. and De Leenheer, A. P. [1996], 'Accuracy-based assessment of proficiency testing results with serum from single donations: possibilities and limitations', *Clinical Chemistry* **42**(3), 469–470.

[161] Strathmann, F. G., Baird, G. S. and Hoffman, N. G. [2011], 'Simulations of delta check rule performance to detect specimen mislabeling using historical laboratory data', *Clinica Chimica Acta* **412**(21-22), 1973–1977.

[162] Streichert, T., Otto, B., Schnabel, C., Nordholt, G., Haddad, M., Maric, M., Petersmann, A., Jung, R. and Wagener, C. [2011], 'Determination of hemolysis thresholds by the use of data loggers in pneumatic tube systems', *Clin Chem* **57**(10), 1390–1397.

[163] Stroobants, A. K., Goldschmidt, H. M. J. and Plebani, M. [2003], 'Error budget calculations in laboratory medicine: linking the concepts of biological variation and allowable medical errors', *Clinica Chimica Acta* **333**(2), 169–176.

[164] Sturgeon, C. M. and Ellis, A. R. [2007], 'Improving the comparability of immunoassays for prostate-specific antigen (PSA): progress and problems', *Clinica Chimica Acta* **381**(1), 85–92.

[165] Sunderman, F. W. [1992], 'The history of proficiency testing/quality control', *Clinical Chemistry* **38**(7), 1205–1209; discussion 1218–1225.

[166] Szecsi, P. B. and Ödum, L. [n.d.], 'Error tracking in a clinical biochemistry laboratory', *Clinical Chemistry and Laboratory Medicine* **47**(10).

[167] Tai, S. S. and Welch, M. J. [2005], 'Development and evaluation of a reference measurement procedure for the determination of estradiol-17β in human serum using Isotope-Dilution liquid Chromatography-Tandem mass spectrometry', *Analytical Chemistry* **77**(19), 6359–6363.

[168] Taieb, J., Benattar, C., Birr, A. S. and Poüs, C. [2003], 'From ACS-180 to Advia-Centaur (Bayer diagnostics): assessment of estradiol, progesterone, LH and FSH assays', *Ann Biol Clin* **61**, 223–228.

[169] Taylor, A. [2011], 'Quality assessment of measurement', *Journal of Trace Elements in Medicine and Biology: Organ of the Society for Minerals and Trace Elements (GMS)* **25 Suppl 1**, S17–21.

[170] Telisman, S., Colak, B., Pizent, A., Jurasovic, J. and Cvitkovic, P. [2007], 'Reproductive toxicity of low-level lead exposure in men', *Environmental Research* **105**(2), 256–266.

[171] Tello, F. L. and Hernández, D. M. [2000], 'Performance evaluation of nine hormone assays on the immulite 2000® Immunoassay system', *Clinical Chemistry and Laboratory Medicine* **38**(10), 1039–1042.

[172] Thienpont, L., Franzini, C., Kratochvila, J., Middle, J., Ricos, C., Siekmann, L. and Stöckl, D. [1995], 'Analytical quality specifications for reference methods and operating specifications for networks of reference laboratories', *European journal of clinical chemistry and clinical biochemistry* **33**(12), 949.

[173] Thienpont, L. M., Brabandere, V. I. D., Stöckl, D. and Leenheer, A. P. D. [1994], 'Use of cyclodextrins for prepurification of progesterone and testosterone from human serum prior to determination with isotope dilution gas chromatography/mass spectrometry', *Analytical chemistry* **66**(22), 4116–4119.

[174] Thienpont, L. M., Nieuwenhove, B. V., Stöckl, D., Reinauer, H. and Leenheer, A. P. D. [1996], 'Determination of reference method values by isotope dilution-gas chromatography/mass spectrometry: a five years' experience of two European Reference Laboratories', *European Journal of Clinical Chemistry and Clinical Biochemistry* **34**(10), 853–860.

[175] Thienpont, L. M., Stöckl, D., Friedecký, B., Kratochvíla, J. and Budina, M. [2003], 'Trueness verification in European external quality assessment schemes: time to care about the quality of the samples', *Scandinavian Journal of Clinical and Laboratory Investigation* **63**(3), 195–201.

[176] Thienpont, L. M., Uytfanghe, K. V., Marriott, J., Stokes, P., Siekmann, L., Kessler, A., Bunk, D. and Tai, S. [2005], 'Feasibility study of the use of frozen human sera in split-sample comparison of immunoassays with candidate reference measurement procedures for total thyroxine and total triiodothyronine measurements', *Clinical Chemistry* **51**(12), 2303–2311.

[177] Tholen, D. W. [2004], 'Impact of international standards and initiatives on proficiency testing for medical laboratories', *Accreditation and Quality Assurance* **9**(11), 653–656.

[178] Thompson, M., Ellison, S. L. R. and Wood, R. [2006], 'The international harmonized protocol for the proficiency testing of analytical chem-

istry laboratories (IUPAC technical report)', *Pure and Applied Chemistry* **78**(1), 145–196.

[179] van den Besselaar, A. M. H. P., Haas, F. J. L. M., van der Graaf, F. and Kuypers, A. W. H. M. [2009], 'Harmonization of fibrinogen assay results: study within the framework of the Dutch project 'Calibration 2000'', *International Journal of Laboratory Hematology* **31**(5), 513–520.

[180] Vesper, H. W., Miller, W. G. and Myers, G. L. [2007], 'Reference materials and commutability', *The Clinical Biochemist Reviews* **28**(4), 139–147.

[181] Vesper, H. W. and Thienpont, L. M. [2009], 'Traceability in laboratory medicine', *Clinical chemistry* **55**(6), 1067.

[182] Waugh, J. M., Collier, C. P., Day, A. G., Waugh, M. and Raymond, M. J. [2002], 'Proficiency testing performance: a case study with modeling', *Clinical Biochemistry* **35**(6), 447–453.

[183] Westgard, J. O. [1994], 'Selecting appropriate quality-control rules', *Clinical Chemistry* **40**(3), 499–501.

[184] Westgard, J. O. [2003], 'Internal quality control: planning and implementation strategies', *Ann Clin Biochem* **40**(6), 593–611.

[185] Westgard, J. O. and Westgard, S. A. [2006*a*], 'The quality of laboratory testing today', *American journal of clinical pathology* **125**(3), 343.

[186] Westgard, J. O. and Westgard, S. A. [2006*b*], 'The quality of laboratory testing today', *American journal of clinical pathology* **125**(3), 343.

[187] White, P. [2004], 'Questionable results-should EQA results be accepted as submitted or should 'obvious mistakes' be corrected? reply to wood WG. questionable results-who directs the EQAS organisers? CCLM 2004; 42 (9): 1073.', *Clinical chemistry and laboratory medicine* **42**(12), 1456.

[188] Whitehead, T. P., Browning, D. M. and Gregory, A. [1973], 'A comparative survey of the results of analyses of blood serum in clinical chemistry laboratories in the united kingdom', *Journal of Clinical Pathology* **26**(6), 435 –445.

[189] Willems, G., Pison, G., Rousseeuw, P. J. and Van Aelst, S. [2002], 'A robust hotelling test', *Metrika* **55**(1-2), 125–138.

[190] Willems, H. L. [2004], 'Wood WG. questionable results–who directs the EQAS organisers?', *Clinical Chemistry and Laboratory Medicine* **42**(9), 1078–1078.

[191] Wilrich, P. [2007], 'Robust estimates of the theoretical standard deviation to be used in interlaboratory precision experiments', *Accreditation and Quality Assurance* **12**(5), 231–240.

[192] Wilson, D. H., Groskopf, W., Hsu, S., Caplan, D., Langner, T., Baumann, M., DeManno, D., Williams, G., Payette, D., Dagel, C., Lynch, D. and Manderino, G. [1998], 'Rapid, automated assay for progesterone on the abbott AxSYM(TM) analyzer', *Clin Chem* **44**(1), 86–91.

[193] Wood, W. G. [2004], 'Questionable results–who directs the EQAS organisers?', *Clinical Chemistry and Laboratory Medicine* **42**(9), 1073–1073.

[194] Working group 1 of the JCGM [1993], 'Evaluation of measurement data - guide to the expression of uncertainty in measurement (JCGM 100:2008)', *www.bipm.org* .

[195] Yang, D. T., Owen, W. E., Ramsay, C. S., Xie, H. and Roberts, W. L. [2004], 'Performance characteristics of eight estradiol immunoassays', *American Journal of Clinical Pathology* **122**(3), 332–337.

[196] Yoshida, H., Imafuku, Y. and Nagai, T. [2004], 'Matrix effects in clinical immunoassays and the effect of preheating and cooling analytical samples', *Clinical Chemistry and Laboratory Medicine* **42**(1), 51–56.

[197] Zaninotto, M., Mion, M. M., Altinier, S., Varagnolo, M., Venturini, R. and Plebani, M. [2009], ‘Performance characteristics of laboratory testing and clinical outcomes’, *Clinica Chimica Acta* **404**(1), 41–45.