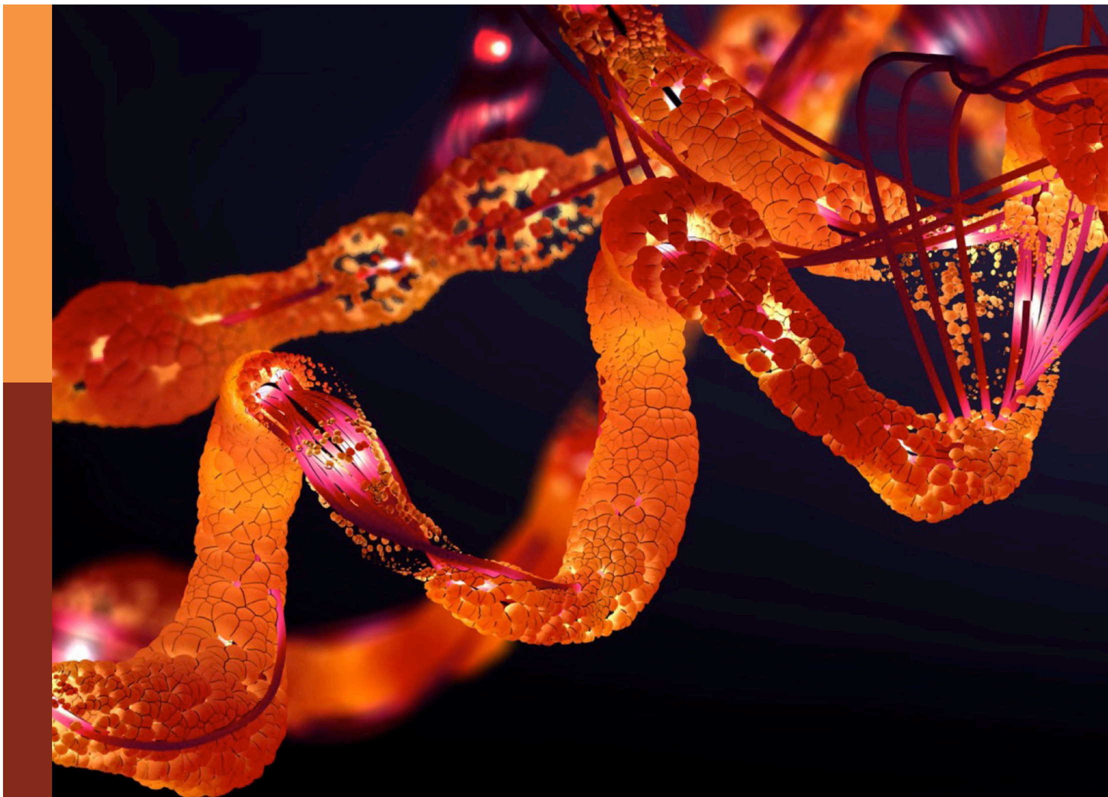
frontiers | Research Topics

Lessons from external quality control in laboratory medicine: Important implications for public health!

Edited by
Klaus-Peter Hunfeld, Hansotto Reiber, Piet Meijer, Peter Luppa,
Dirk Schlüter, Michael Spannagl, Douglas Norris and Ingo Schellenberg

Published in
Frontiers in Molecular Biosciences
Frontiers in Immunology



ISSN 1664-8714
ISBN 978-2-8325-5472-2
DOI 10.3389/978-2-8325-5472-2

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Lessons from external quality control in laboratory medicine: Important implications for public health!

Topic editors

Klaus-Peter Hunfeld — Goethe University, Germany
Hansotto Reiber — University Medical Center Göttingen, Germany
Piet Meijer — ECAT Foundation, Netherlands
Peter Luppa — Technical University of Munich, Germany
Dirk Schlüter — Hannover Medical School, Germany
Michael Spannagl — Ludwig Maximilian University of Munich, Germany
Douglas Norris — Johns Hopkins University, United States
Ingo Schellenberg — INSTAND e.V., Germany

Citation

Hunfeld, K.-P., Reiber, H., Meijer, P., Luppa, P., Schlüter, D., Spannagl, M., Norris, D., Schellenberg, I., eds. (2024). *Lessons from external quality control in laboratory medicine: Important implications for public health!* Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5472-2

Table of contents

TYPE Editorial
PUBLISHED 12 September 2024
DOI 10.3389/fmolb.2024.1485193

OPEN ACCESS

EDITED AND REVIEWED BY
Matteo Becatti,
University of Firenze, Italy

*CORRESPONDENCE
Klaus-Peter Hunfeld,
hunfeld.klaus-peter@khnw.de

RECEIVED 23 August 2024
ACCEPTED 23 August 2024
PUBLISHED 12 September 2024

CITATION
Weiss N, Schellenberg I and Hunfeld K-P (2024) Editorial: Lessons from external quality control in laboratory medicine: important implications for public health! *Front. Mol. Biosci.* 11:1485193.
doi: 10.3389/fmolb.2024.1485193

Editorial: Lessons from external quality control in laboratory medicine: important implications for public health!

Nathalie Weiss[1], Ingo Schellenberg[1,2] and Klaus-Peter Hunfeld[1,3]*

[1]INSTAND e.V., Society for Promoting Quality Assurance in Medical Laboratories e.V., Duesseldorf, Germany, [2]Center of Life Sciences, Institute of Bioanalytical Sciences (IBAS), Anhalt University of Applied Sciences, Bernburg, Germany, [3]Northwest Medical Centre, Academic Teaching Hospital, Medical Faculty, Institute for Laboratory Medicine, Microbiology & Infection Control, Goethe University Frankfurt, Frankfurt, Germany

KEYWORDS

proficiency testing (PT), external quality assessment (EQA), public health, laboratory medicine, longitudinal analysis

Editorial on the Research Topic
Lessons from external quality control in laboratory medicine: important implications for public health!

1 Introduction

Laboratory medicine has gone a long way from the optical and olfactory analysis of urine in the 16th century to modern day laboratory diagnostics, where even a small amount of nucleic acid from human pathogens can be detected out of a matrix of a variety of other molecules. Moreover, currently, up to 80% of medical decision making is supported by laboratory analysis, highlighting the substantial impact of laboratory medicine on public health (Salinas, 2023). The need for reliable laboratory results, regardless of testing site or method, gained international attention when a survey by Belk and Sunderman showed that 42% of laboratory results for glucose and 49% for hemoglobin were of insufficient quality (Belk et al., 1947). Then, as now, those figures were unacceptable and posed a clear threat to public health. This publication is often cited as the kick-off for the success story of external quality assessment (EQA) schemes, also known as proficiency testing (Doxiadis et al., 2024). Today, EQA schemes are a mandatory part of various national and international guidelines like the Clinical Laboratory Improvement Amendments of 1988 (CLIA) of the United States of America (USA) (Clinical laboratory improvement Amendments of 1988) or the Guideline of the German Medical Association (Richtlinie der Bundesärztekammer Rili-BÄK) (Bundesärztekammer, 2023). In addition, laboratories seeking accreditation, e.g., according to ISO 15189, are obliged to participate in external quality assurance procedures (ISO, 2023).

The organization of EQA schemes is mostly done by non-profit organizations like INSTAND e.V., CAP or UK-NEQAS and some commercial institutions like BioRad and all institutions are accredited according to ISO 17043. Although the guideline provides a good framework for the organization of high-quality EQA schemes, EQA providers still face some hurdles such as the provision of suitable samples for the quality assessment. One intensely discussed Research Topic is the commutability of EQA samples, meaning their exchangeability with patient samples. In their opinion paper, Vierbaum et al. address the currently published models of commutability assessment and discuss them in the context of feasibility particularly for the providers of EQA schemes.

This feasibility is especially challenged when new analytes or methods are introduced to the *in vitro* diagnostic market. A growing market is focused on point-of-care testing (POCT, also known as bedside diagnostics) and the analytical increase in this area of laboratory medicine has also been observed by Luppa et al. for the detection of glucose with a rise in participating laboratories in the POCT glucose EQA program in Germany. In addition, they put the EQA results for HbA1c and POCT-glucose in perspective to the current diagnostic methodology of tests for diabetes as well as morbidity and mortality of diabetes patients, showing that the quality of the measurement of HbA1c clearly improved over time.

The positive impact of accreditation status and analytical methods on the likelihood of satisfactory results in the detection of *Escherichia coli* in environmental samples by Canadian environmental testing laboratories was demonstrated by Sreya et al. and the authors propose an implementation of regulated EQA schemes in drinking water safety plans.

New clinical variants of bacteria challenge laboratories and physicians alike, and therefore it is important to include these variants in quality assessments for training and for testing the quality of the assays used. In support, Kremser et al. showed a positive effect of using new clinical variants for EHEC/STEC, *B. burgdorferi* and MRSA/cMRSA in their longitudinal evaluation of EQA schemes for the detection of these bacteria using nucleic acid amplification techniques (NAAT).

Bacteria pose a growing threat to public health due to increasing antimicrobial resistance (AMR) (Antimicrobial Resistance Collaborators, 2022; The Lancet, 2024). Therefore, the correct identification and subsequent susceptibility testing of bacteria is of paramount importance for therapy and care of infectious diseases. Here, Lindenberg et al. examined 17 years of EQA schemes for bacterial identification and susceptibility testing and showed that while the quality of bacterial identification remained consistently high, the quality of AMR testing was affected by laboratory type as well as changes in testing guidelines and unregulated adherence to these guidelines.

The recent COVID-19 pandemic and the Mpox outbreaks have brought the detection of viral genetic material by NAAT into the focus of specialists and manufacturers. In particular, new NAAT methods for whole genome (WG) sequencing and the subsequent encouragement by authorities to use this method have raised questions about the current quality of these methods. Interestingly, Camp et al. identified hurdles in building next-generation sequencing capacity in diagnostic laboratories in Austria, but the overall quality of analyses was good, with a few exceptions that clearly showed improvement in quality over time.

While the pandemic shifted the focus from classical laboratory medicine more to the detection of infectious diseases, other analytes gained also in importance. Accordingly, Kirschfink et al. observed an increasing interest in complement EQA programs since 2016. While the pass rates for C3, C4, C1 inhibitor antigen and activity determinations provided good proficiency testing results, the activation pathways showed greater variance, especially for pathological samples, highlighting the need for further improvement and harmonization.

However, there is also room for improvement for well-established biomarkers such as high-sensitivity (hs)-CRP. Weiss et al. were able to show that there is a positive trend towards harmonization, based on EQA data. However, the persistence of manufacturer-specific differences underlines the further need for meta-analytics stratified by assay in order to gain a meaningful insight into the usefulness of this marker for cardiovascular risk assessment. Similarly, the analyses by Toll et al. show the interesting evolution of the EPO EQA program, that was first introduced in 2017, from its first steps till now. It highlights the difficulties and opportunities of new EQA surveys for blood and serum markers and the potential impact of reference materials and methods is discussed further. Kremser et al. also emphasized the need for international reference materials based on their longitudinal evaluation of EQA schemes for cancer antigen tumor markers. The methods used by the different laboratories showed high precision within methods but considerable variability between methods, underlining the fact that the same patient should only be monitored with the same method.

For EQAs the meaningful analysis and interpretation of statistical data is extremely important. While most EQA programs can only be evaluated based on the consensus mean or results from expert laboratories due to the lack of a reference method and reference material, these are established for some markers such as steroid hormones. In their study, Vierbaum et al. were able to determine the accuracy of several immunoassays for the detection of testosterone, progesterone and 17β-estradiol in serum based on EQA data. Whereas improvement in standardization is required for accurate analysis and thus clinically reliable interpretations, one manufacturer showed increasing accuracy over the observed time period.

Unfortunately, the statistical analysis of analytes in biological matrices such as blood or urine is not as simple, as the values are not guaranteed to follow a normal distribution. Seifert et al. propose a logit transformation for the analysis of data points near 0% or 100% to generate a symmetric distribution with zero center, so that parametric statistical methods can be used without bias.

And while generating reliable laboratory results is important for medical diagnosis, one important factor should never be overlooked: The patient is more than just a measurable analyte. In a medico philosophical article Reiber's hypothesis and theories are discussed to promote the need and opportunity for CSF diagnostic reports that integrate all patient data rather than looking at individual markers, which serves better care for the individual patient but can also help reducing costs for the healthcare system as a whole.

2 Conclusion and perspective

Overall, this Research Topic on external quality control in laboratory medicine provides an excellent overview of the impact and importance of proficiency testing in different areas of laboratory diagnostics and public health. The different authors from various areas of laboratory medicine have not only provided a well-rounded picture of proficiency testing tools and the impact of the obtained results on public health and quality of care but, most importantly, have also shared their critical thoughts and opinions on current principles and future opportunities for EQA in the years to come.

Author contributions

NW: Writing–original draft, Writing–review and editing. IS: Writing–original draft, Writing–review and editing. K-PH: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We are grateful to all the authors and reviewers for their enormous contributions to this Research Topic.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were editorial board members of Frontiers, at the time of submission. This had no impact on the peer review process and the final decisions.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Antimicrobial Resistance Collaborators (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*. 399(10325), 629–655. doi:10.1016/S0140-6736(21)02724-0

Belk, W. P., and Sunderman, F. W. (1947). A survey of the accuracy of chemical analyses in clinical laboratories. *Am. J. Clin. Pathol.* 17(11), 853–861. doi:10.1093/ajcp/17.11.853

Bundesärztekammer (2023). Richtlinie der Bundesärztekammer zur Qualitätssicherung laboratoriumsmedizinischer Untersuchungen. Deutschen Ärzteblatt. 120. Available at: https://www.bundesaerztekammer.de/fileadmin/user_upload/BAEK/Themen/Qualitaetssicherung/_Bek_BAEK_RiLi_BAEK_ONLINE_FINAL_VERS_26_05_2023.pdf.

Clinical laboratory improvement Amendments of 1988 (CLIA), Pub. L. No 100 -578, 102 Stat. 2903 (1988). (Accessed April 29, 2024).

Doxiadis, I., and Lehmann, C. (2024). External proficiency testing exercises: challenges and opportunities. *Front. Genet.* 15, 1304312. doi:10.3389/fgene.2024.1304312

ISO (2023). Medical laboratories—Requirements for quality and competence (ISO 15189:2022). German version EN ISO 15189:2022.

The Lancet (2024). Antimicrobial resistance: an agenda for all. *Lancet*. 403(10442), 2349. doi:10.1016/S0140-6736(24)01076-6

Salinas, M. (2023). Laboratory Medicine: from just testing to saving lives. *Clin. Chem. Lab. Med.* 61(10), 1677–1678. doi:10.1515/cclm-2023-0379

frontiers | Frontiers in Medicine

TYPE Opinion
PUBLISHED 01 August 2024
DOI 10.3389/fmed.2024.1416642

OPEN ACCESS

EDITED BY
Hem Chandra Jha,
Indian Institute of Technology Indore, India

REVIEWED BY
Gunnar Nordin,
Equalis AB, Sweden
Anna Frances Lau,
National Institutes of Health (NIH),
United States

*CORRESPONDENCE
Laura Vierbaum
✉ vierbaum@instand-ev.de

RECEIVED 24 April 2024
ACCEPTED 22 July 2024
PUBLISHED 01 August 2024

CITATION
Vierbaum L, Kaiser P, Spannagl M, Wenzel F, Thevis M and Schellenberg I (2024) Experiences and challenges for EQA providers in assessing the commutability of control materials in accuracy-based EQA programs. *Front. Med.* 11:1416642. doi: 10.3389/fmed.2024.1416642

Experiences and challenges for EQA providers in assessing the commutability of control materials in accuracy-based EQA programs

Laura Vierbaum[1]*, Patricia Kaiser[1], Michael Spannagl[1,2], Folker Wenzel[1,3], Mario Thevis[4] and Ingo Schellenberg[1,5]

[1]INSTAND e.V., Society for Promoting Quality Assurance in Medical Laboratories, Duesseldorf, Germany, [2]Institute of Laboratory Medicine, Ludwig-Maximilians-University Munich, Munich, Germany, [3]Faculty of Medical and Life Sciences, Furtwangen University, Villingen-Schwenningen, Germany, [4]Institute of Biochemistry/Center for Preventive Doping Research, German Sport University Cologne, Cologne, Germany, [5]Institute of Bioanalytical Sciences (IBAS), Center of Life Sciences, Anhalt University of Applied Sciences, Bernburg, Germany

KEYWORDS

external quality assessment (EQA), proficiency testing (PT), commutability, reference measurement procedures, trueness control, traceability, measurement accuracy

Introduction

External quality assessment (EQA) programs in medical laboratory diagnostics are necessary for gaining insight into the analytical performance of a large number of analytes across laboratories. The evaluation of EQA results is either based on consensus values of individual measurement procedures (MPs) or is accuracy-based using a reference measurement value (RMV) as target value for all MPs if a reference MP (RMP) has been established for the analyte. RMPs achieve the highest possible analytical accuracy as they are, ideally, matrix-independent and are comparatively insensitive to interfering substances. In contrast, MPs of a lower metrological order, as are commonly used for routine diagnostics in medical laboratories, can be affected to varying degrees by the sample matrix and components. That is depending on the analytical performance of the respective MP.

Thus, if an RMV can be assigned as a target value to the EQA material (EQAM), EQA programs can serve as measurement trueness controls (1). However, this presupposes that the EQAMs are suitable for measurement with all MPs (2). Since control materials (CMs) are usually processed, their exchangeability with patient samples, also known as commutability (3–5), should be aimed for and is increasingly being called for in professional circles. Due to the required use of pathological analyte concentrations, and the need for samples for up to several hundred participating laboratories, the use of native single donor samples in EQA schemes is not feasible from an ethical point of view, and pooling and spiking of CMs is common. In addition, CMs are commonly stabilized by means of stabilizing additives or lyophilization. While the RMPs used to assign target values ideally remain unaffected by the processed nature of EQAMs, individual routine MPs might be affected, resulting in artificial shifts in MP-specific bias. MP-specific effects on the analysis due to the processing of the material tend to be more critical for analytes with a more complex or possibly tertiary structure, such as proteins (6, 7), than for lower molecular mass analytes with a simpler structure, such as urea. Thus, if observed bias are in part due to a lack of commutability of the EQAM, an EQA evaluation is only possible within the MP collectives based on the consensus value and not on the level of accuracy.

Hence, it is clear that the use of commutable CMs is beneficial to improve quality assurance in medical laboratories. However, the investigation of commutability poses enormous challenges for EQA providers worldwide and requires elaborate practice-oriented concepts.

Models for commutability assessments

Commutability assessment approaches were originally developed in the field of clinical chemistry, where RMPs are established for some analytes and higher order reference standards exist. However, possible influences of matrix effects on the measurement results for CMs are relevant in all disciplines, e.g., hematology, immunology or virology, and the statistical models for assessing the commutability of CMs described in guidelines can be transferred to analytes in other fields to the extent possible.

It is recommended measuring the analyte in at least 30 native patient samples and measuring CMs using as many MPs on the market as possible. The measurement results of two MPs in any combination can then be correlated, recommended as Deming regression (3), or as a difference in bias plot, also known as a Bland-Altman plot (4, 5). If an RMP is available for the analyte in question, the measurement results of all MPs can only be correlated with the RMVs. A range is then defined based on the correlations of the patient sample values and used as a commutability criterion for the CMs.

The relevance of such models for assessing commutability of processed CMs is beyond question, but the challenges involved in the practical implementation of such theoretical models are virtually impossible for EQA providers to realize. Consequently, limitations prevail that necessitate consideration and awareness as discussed in the following.

Recruitment of donors for commutability assessment and limitations

To obtain correlations of the measured values that are representative of patient samples, these should cover the broadest possible concentration range of the respective analytes. Recruiting 30 donor materials takes great effort and pushes the limits of what is feasible. Firstly, the targeted selection of suitable donors requires a pre-characterization of numerous patients. Secondly, and most critically, the donation of samples in the range of pathological values is ethically debatable. But MPs can deliver conspicuous results, especially in high or low concentration ranges (8, 9), so that including such patient samples in commutability assessments might be crucial. To avoid freezing patient samples, which can cause, for example, changes in protein structures, all donations for commutability assessment must be collected and processed on the same day and measured immediately by as many MPs as possible. But some analytes have a high biological variability, so a pre-characterization of patients at a certain time would not assist in patient selection to cover the desirable broad concentration range with the fresh specimens. For example, blood parameters, like glucose or electrolytes, fluctuate substantially depending on food intake or fluid balance (10).

Definition of acceptability criteria and limitations

Ultimately, it is questionable whether even 30 patient samples are sufficient to represent the diverse patient profile and to make a fundamental statement on the commutability of CMs on this basis. Moreover, these models for commutability assessment do not take into account the fact that patient samples may also contain interfering substances for individual MPs, especially those from ill or medicated patients. The fundamental question that arises when we assess commutability is what are suitable criteria for selecting reference patient samples.

While the criteria for commutability of CMs might be narrowly defined based on exemplary patient samples, possible MP weaknesses may be missed. The IFCC Working Group states that MPs with inadequate precision are not suitable for assessing commutability with the Bland-Altman plot, as this might impact the assessment (4). When using the Bland-Altman plot with medically-diagnostically defined criteria, patient samples might also appear to be non-commutable if inadequately precise MPs are included in the assessment. However, when assessing a material, where should the line be drawn between what is considered to be adequate and inadequate in terms of an MP's precision? The line is blurred between whether measurement differences are caused by inadequate precision or material effects.

The assessment of commutability of CMs, of course, depends largely on the strictness of the criteria. The commutability criteria in the Bland-Altman plot can be defined and subjectively justified in different ways. Tight criteria can lead to inconclusive results as the values of the CMs including the measurement uncertainties must fit the criteria (6). Due to this leeway in defining the criteria for the Bland-Altman plot on the one hand, and, the statistically defined criteria for the Deming regression on the other, it is not surprising that an application of both models for one and the same data set yields different results (11).

The role of measurement performance of MPs

The evaluation of EQA results based on RMVs can reveal a bias in an MP and insufficient standardization of a diagnostic test system. Biased results can indicate an MP's lack of accuracy, yet bias can also be caused by an EQAM's lack of commutability. However, it is hard to identify the contributions of these two parties to an observed measurement bias. It should be noted that the analytical selectivity of an MP to interferences also determines whether it is affected by matrix effects or sample additives. Analyses of data from past EQA surveys show that MPs of the market-leading manufacturers deliver measurements with varying degrees of robustness and accuracy. However, individual MPs manage to reliably deliver very precise and some also very accurate results (12, 13), even when measurements are conducted on EQAMs of a processed nature.

Lack of specificity or low robustness of certain MPs have also been identified in studies with clinical samples and are the core reason behind unreliable laboratory diagnostics (14, 15). Measurement results should be reliable, especially for "conspicuous" patient samples, e.g., for samples from patients under the influence of medication, where undesirable disturbances in the analysis can occur more frequently (16).

The measurement of creatinine is an example of a clearly divided distribution of INSTAND EQA results depending on whether kinetic or enzymatic methods were used for the analysis.[1] The results from interlaboratory testing, classified as a category 1 EQA scheme, on samples assessed as commutable (17) show that serum creatinine measurements were more accurate using enzymatic methods than kinetic ones (18–20). The Jaffe method reagent is known to be sensitive toward reacting with several interfering components in serum such as glucose, bilirubin, or hemoglobin, which is consequently critical for measurements on icteric or hemolytic samples. Thus, a lack of specificity in the kinetic creatinine measurement produces overestimated values, especially in the case of lower creatinine concentrations (21, 22).

For other measurands in clinical chemistry, there are no means of metrological traceability for the values. For example, there are no high-order MPs or primary standards for procalcitonin measurement (23). In such circumstances, a high variability in EQA results is not surprising and an evaluation of the analytical performance of a laboratory can only be made based on consensus values.

Challenges for EQA providers

Commutability assessment studies are quite feasible for certified reference materials that are usually produced on a large scale. However, the situation is different for EQA providers who offer EQA schemes for many analytes in laboratory medicine and who manage a high throughput of batches per scheme and year. Consequently, an enormous number of studies would be necessary to investigate the commutability of the high number of different EQAM batches. With the requirement for commutable CMs, EQA providers are challenged by what is feasible and financially viable. The high number of patient samples required for material assessment, even in pathological concentration ranges, appears very paradoxical and practically impossible to implement when one considers that processed materials are deliberately used in EQA schemes, particularly for ethical reasons. Severely ill patients in the areas of hematology, immunohematology, and oncology, including those undergoing therapy, cannot have large quantities of blood taken so that EQA schemes can be conducted with several hundred participating laboratories or numerous commutability studies can be performed to characterize the EQAMs. Also for ethical reasons, samples from patients with rare diseases cannot be included in such surveys. The availability of patient materials for studies is also severely limited if the collection of the material is associated with increased medical intervention, as in the case of cerebrospinal fluid. The applicability of the models for assessing commutability is thus severely limited to clinical chemistry parameters.

Commutability assessments of EQAMs is far from feasible for MPs that are less prevalent on the market, and especially for in-house products. EQA providers can only include the market-leading MPs in commutability studies in cooperation with representative measurement centers.

In order to significantly reduce the effort for EQA providers to provide commutability studies for all EQAM batches, it is sometimes assumed that the results of a study can also be applied to identically produced CMs. However, it is known that lot-to-lot variability can occur even with identically produced sample or assay batches (24). Occasional effects in EQAM batches are represented by a conspicuous and unusual scattering in the MP-specific value distribution, e.g., as observed in 2022 in an INSTAND EQA scheme for the quantitation of 17β-estradiol (12). Samples with such conspicuous results must be excluded from the EQA evaluation.

Overall, the effort needed for the commutability studies is a challenge that cannot currently be overcome in practice. EQA providers can only check in bullet points and to a limited extent the commutability of the EQAMs. Due to limitations in the implementation of commutability studies, which include the lack of availability of patient samples with pathological concentrations, a focus on market-leading methods, and varying criteria definitions, the information obtained from these studies should be balanced against the enormous effort that they entail. Obtaining a broader and more comprehensive picture is certainly scientifically desirable, however, the inevitable limitations of the assessments, costs, and EQA fees become incompatible with client needs. Hence, it appears that the only way to achieve progress regarding the challenge of assessing the commutability of EQAMs in a practical manner necessitates the cooperation of sample manufacturers, national metrology institutes, and IVD manufacturers and cannot be handled by EQA providers alone.

Ultimately it is the aim of EQA providers to offer the most suitable CMs possible to support reliable measurement results in medical laboratories. Comparative studies with patient materials on a smaller scale and empirical values from the literature can provide valuable indications whether significant sample effects are expected for an EQAM.

Relevance of the exact study design

The widespread dissemination of the theoretical commutability assessment models in medical and scientific communities has led to increased demand for commutable CMs. In reaction to this, CMs are increasingly being declared commutable without providing more detailed information on the study design.

Commutability is not a property that can be attributed exclusively to the material but must be regarded as being a direct result of the exact study design (25). At the very least, information needs to be provided about the MPs and the defined assessment criterion involved in order to gain an accurate picture of a material's commutability. The complexity of commutability assessments is often not considered in its entirety in professional circles and by CM end users. In light of the fact that the implementation of commutability studies necessitates major and minor limitations,

[1] INSTAND *RV-Online* [Online]. Available online at: http://rv-online.instandev.local/index.shtml?lang=en.

uncommented statements on the commutability of CMs should always be interpreted critically.

Conclusion

The aim of EQA providers is to promote quality assurance in medical laboratories. Thus, it is in their interest to provide EQAMs that are as suitable as possible for the purpose of the EQA. However, the models for commutability assessment of CMs are only theoretical models that reach their limits in practice in terms of their practicability. Assessing all EQAM batches is simply not manageable and sporadic assessments come up against several limitations. In particular, the availability of patient samples with pathological concentrations is critical and a focus on market-leading MPs is necessary. Statements on commutability must necessarily be interpreted within the context of the entire study design. Ultimately, the information gained from these assessments should be balanced against the enormous effort involved, and practice-oriented concepts need to be developed, which would greatly benefit from the cooperation between all parties involved, EQA providers, sample manufacturers, national metrology institutes, and IVD manufactures in internationally active networks.

Author contributions

LV: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. PK: Supervision, Writing – review & editing, Conceptualization, Methodology. MS: Supervision, Writing – review & editing, Methodology. FW: Writing – review & editing, Supervision. MT: Supervision, Writing – review & editing. IS: Methodology, Supervision, Writing – review & editing, Conceptualization, Project administration.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. ISO/IEC Guide 99. *International vocabulary of metrology – Basic and general concepts and associated terms (VIM).* ISO/IEC Guide 99:2007. (2007).

2. Miller WG, Jones GR, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. *Clin Chem.* (2011) 57:1670–80. doi: 10.1373/clinchem.2011.168641

3. CLSI. *Characterization and Qualification of Commutable Reference Materials for Laboratory Medicine; Approved Guideline. CLSI document EP30-A.* Wayne, PA: Clinical and Laboratory Standards Institute (2010).

4. Miller WG, Schimmel H, Rej R, Greenberg N, Ceriotti F, Burns C, et al. IFCC working group recommendations for assessing commutability part 1: general experimental design. *Clin Chem.* (2018) 64:447–54. doi: 10.1373/clinchem.2017.277525

5. Nilsson G, Budd JR, Greenberg N, Delatour V, Rej R, Panteghini M, et al. IFCC working group recommendations for assessing commutability part 2: using the difference in bias between a reference material and clinical samples. *Clin Chem.* (2018) 64:455–64. doi: 10.1373/clinchem.2017.277541

6. Delatour V, Clouet-Foraison N, Jaisson S, Kaiser P, Gillery P. Trueness assessment of HbA1c routine assays: are processed EQA materials up to the job? *Clin Chem Lab Med.* (2019) 57:1623–31. doi: 10.1515/cclm-2019-0219

7. Dikaios I, Althaus H, Angles-Cano E, Ceglarek U, Coassin S, Cobbaert CM, et al. Commutability assessment of candidate reference materials for lipoprotein(a) by comparison of a MS-based candidate reference measurement procedure with immunoassays. *Clin Chem.* (2023) 69:262–72. doi: 10.1093/clinchem/hvac203

8. La'ulu SL, Kalp KJ, Straseski JA. How low can you go? Analytical performance of five automated testosterone immunoassays. *Clin Biochem.* (2018) 58:64–71. doi: 10.1016/j.clinbiochem.2018.05.008

9. Ward G, Simpson A, Boscato L, Hickman PE. The investigation of interferences in immunoassay. *Clin Biochem.* (2017) 50:1306–11. doi: 10.1016/j.clinbiochem.2017.08.015

10. Thomas L. *Electrolyte and Water Balance.* Clinical Laboratory Diagnostics (2020).

11. Yue Y, Zhang S, Xu Z, Chen X, Wang Q. Commutability of reference materials for α-fetoprotein in human serum. *Arch Pathol Labor Med.* (2017) 141:1421–7. doi: 10.5858/arpa.2016-0441-OA

12. Vierbaum L, Weiss N, Kaiser P, Kremser M, Wenzel F, Thevis M, et al. Longitudinal analysis of external quality assessment of immunoassay-based steroid hormone measurement indicates potential for improvement in standardization. *Front Molec Biosci.* (2024) 11:1345356. doi: 10.3389/fmolb.2024.1345356

13. Wojtalewicz N, Vierbaum L, Kaufmann A, Schellenberg I, Holdenrieder S. Longitudinal evaluation of AFP and CEA external proficiency testing reveals need for method harmonization. *Diagnostics.* (2023) 13:2019. doi: 10.3390/diagnostics13122019

14. Coucke W, Devleeschouwer N, Libeer JC, Schiettecatte J, Martin M, Smitz J. Accuracy and reproducibility of automated estradiol-17beta and progesterone assays using native serum samples: results obtained in the Belgian external assessment scheme. *Hum Reprod.* (2007) 22:3204–9. doi: 10.1093/humrep/dem322

15. Infusino I, Frusciante E, Braga F, Panteghini M. Progress and impact of enzyme measurement standardization. *Clin Chem Lab Med.* (2017) 53:334–40. doi: 10.1515/cclm-2016-0661

16. Bojko L, Ripka GP, Dionísio LM, Borges CL, Borato DCK, Moss MF. Drug dosing using estimated glomerular filtration rate: misclassification due to metamizole interference in a creatinine assay. *Ann Clin Biochem.* (2021) 58:474–80. doi: 10.1177/00045632211020029

17. Miller WG, Myers GL, Gantzer ML, Kahn SE, Schönbrunner ER, Thienpont LM, et al. Roadmap for harmonization of clinical laboratory measurement procedures. *Clin Chem.* (2011) 57:1108–17. doi: 10.1373/clinchem.2011.164012

18. González-Lao E, Díaz-Garzón J, Corte Z, Ricós C, Perich C, Álvarez V, et al. Category 1 external quality assessment program for serum creatinine. *Ann Transl Med.* (2017) 5:133. doi: 10.21037/atm.2017.03.70

19. Jeong TD, Cho EJ, Lee K, Lee W, Yun YM, Chun S, et al. Recent trends in creatinine assays in Korea: long-term accuracy-based proficiency testing survey data by the korean association of external quality assessment service (2011-2019). *Ann Lab Med.* (2021) 41:372–9. doi: 10.3343/alm.2021.41.4.372

20. Ricós C, Perich C, Boned B, González-Lao E, Diaz-Garzón J, Ventura M, et al. Standardization in laboratory medicine: two years' experience from category 1 EQA programs in Spain. *Biochem Med.* (2019) 29:010701. doi: 10.11613/BM.2019.010701

21. Choosongsang P, Bhornsrivathanyou N, Aiadsakun P, Choosongsang P, Bodhikul A, Yamsuwan Y, et al. Glucose interference in serum and urine samples with various creatinine concentrations measured by the jaffe kinetic method. *Ejifcc.* (2023) 34:57–65.

22. Syme NR, Stevens K, Stirling C, McMillan DC, Talwar D. Clinical and analytical impact of moving from jaffe to enzymatic serum creatinine methodology. *J Appl Lab Med.* (2020) 5:631–42. doi: 10.1093/jalm/jfaa053

23. Huynh H-H, Bœuf A, Vinh J, Delatour V, Delatour V, Bœuf A, et al. Evaluation of the necessity and the feasibility of the standardization of procalcitonin measurements: activities of IFCC WG-PCT with involvement of all stakeholders. *Clin Chim Acta.* (2021) 515:111–21. doi: 10.1016/j.cca.2021.01.004

24. Luo Y, Pehrsson M, Langholm L, Karsdal M, Bay-Jensen AC, Sun S. Lot-to-lot variance in immunoassays-causes, consequences, and solutions. *Diagnostics.* (2023) 13:1835. doi: 10.3390/diagnostics13111835

25. Vogeser M, Habler K. Is commutability of a reference material always desirable? *J Mass Spectrom Adv Clin Lab.* (2024) 31:17–8. doi: 10.1016/j.jmsacl.2023.12.002

TYPE Original Research
PUBLISHED 31 January 2024
DOI 10.3389/fmolb.2024.1345356

Longitudinal analysis of external quality assessment of immunoassay-based steroid hormone measurement indicates potential for improvement in standardization

Laura Vierbaum[1]*, Nathalie Weiss[1], Patricia Kaiser[1], Marcel Kremser[1], Folker Wenzel[1,2], Mario Thevis[3], Ingo Schellenberg[1,4] and Peter B. Luppa[1,5]

[1]INSTAND e.V., Society for Promoting Quality Assurance in Medical Laboratories, Duesseldorf, Germany, [2]Faculty of Medical and Life Sciences, Furtwangen University, Villingen-Schwenningen, Germany, [3]Institute of Biochemistry/Center for Preventive Doping Research, German Sport University Cologne, Cologne, Germany, [4]Institute of Bioanalytical Sciences (IBAS), Center of Life Sciences, Anhalt University of Applied Sciences, Bernburg, Germany, [5]Institute of Clinical Chemistry and Pathobiochemistry, University Hospital Rechts der Isar, Technische Universität München, Munich, Germany

OPEN ACCESS

EDITED BY
Hem Chandra Jha,
Indian Institute of Technology Indore, India

REVIEWED BY
Julie Brossaud,
Université de Bordeaux, France
Sarah Gregory,
University of Edinburgh, United Kingdom

*CORRESPONDENCE
Laura Vierbaum,
✉ vierbaum@instand-ev.de

RECEIVED 27 November 2023
ACCEPTED 02 January 2024
PUBLISHED 31 January 2024

CITATION
Vierbaum L, Weiss N, Kaiser P, Kremser M, Wenzel F, Thevis M, Schellenberg I and Luppa PB (2024), Longitudinal analysis of external quality assessment of immunoassay-based steroid hormone measurement indicates potential for improvement in standardization. *Front. Mol. Biosci.* 11:1345356.
doi: 10.3389/fmolb.2024.1345356

As hormonal disorders are linked to several diseases, the accurate quantitation of steroid hormone levels in serum is crucial in order to provide patients with a reliable diagnosis. Mass spectrometry-based methods are regarded as having the highest level of specificity and sensitivity. However, immunoassays are more commonly used in routine diagnostics to measure steroid levels as they are more cost effective and straightforward to conduct. This study analyzes the external quality assessment results for the measurement of testosterone, progesterone and 17β-estradiol in serum using immunoassays between early 2020 and May 2022. As reference measurement procedures are available for the three steroid hormones, the manufacturer-specific biases were normalized to the reference measurement values. The manufacturer-specific coefficients of variation were predominantly inconspicuous, below 20% for the three hormones when outliers are disregarded, however there were large differences between the various manufacturer collectives. For some collectives, the median bias to the respective reference measurement value was repeatedly greater than ±35%, which is the acceptance limit defined by the German Medical Association. In the case of testosterone and progesterone determination, some collectives tended to consistently over- or underestimate analyte concentrations compared to the reference measurement value, however, for 17β-estradiol determination, both positive and negative biases were observed. This insufficient level of accuracy suggests that cross-reactivity continues to be a fundamental challenge when antibody detection is used to quantify steroids with a high structural similarity. Distinct improvements in standardization are required to provide accurate analysis and thus, reliable clinical interpretations. The increased accuracy of the AX immunoassay for testosterone measurement, as observed in the INSTAND EQAs between 2020 and 2022, could be the result of a recalibration of the assay and raises hope for further improvement of standardization of immunoassay-based steroid hormone analyses in the coming years.

KEYWORDS

external quality assessment (EQA), proficiency testing (PT), steroid hormones, immunoassays, accuracy, standardization, testosterone, progesterone and 17β-estradiol

1 Introduction

Hormones are biochemical messengers that play a key role in regulating the complex processes of human metabolism. Steroid hormones, such as testosterone, progesterone and 17β-estradiol, control the gender formation and maturation, as well as human reproductive processes.

Steroid hormone disorders are linked to a wide variety of health impairments, e.g., menstrual cycle disorders, puberty disorders, and infertility in men and women caused by hypogonadism (Corona et al., 2011; Skałba and Guz, 2011; Kleine and Rossmanith, 2013; Basaria, 2014; Beneke et al., 2015; Klein et al., 2017). This is often accompanied by mental stress for those affected. Pediatric indications also need to be considered, as many steroid disorders of the adrenal cortex first arise in childhood (Salonia et al., 2019; Yadav and Sharma, 2023). In addition to providing diagnostic results, steroid hormone levels are also measured in serum during fertilization and treatment monitoring (Aubard et al., 1997; Gleicher et al., 2000; Strawn et al., 2000; Zitzmann and Nieschlag, 2000; Diemer et al., 2016; Thomsen et al., 2018; Barbonetti et al., 2020; Armeni et al., 2021). Furthermore, elevated hormone levels in serum can be caused by hormone-producing tumors, both in the adrenal cortex and the gonads (Kleine and Rossmanith, 2013; Beneke et al., 2015).

The high biological variability in hormone levels, caused, for example, by circadian rhythms, individual daily variability, temporary stressors, and the menstrual cycle (Beneke et al., 2015), makes the accurate and reliable determination of hormone levels even more important for diagnostic purposes and treatment monitoring. Gas chromatography (GC) or liquid chromatography (LC) coupled mass spectrometry (MS) is the most reliable method to quantify hormones and is thus considered the "gold" standard (Krone et al., 2010; Stanczyk and Clarke, 2010; Conklin and Knezevic, 2020). However, the procedure is both costly and time-consuming and requires a highly qualified laboratory staff. Therefore, immunoassays are currently still the primary method used for routine clinical measurements. However, previous studies have found discrepancies in the measured serum concentrations of sex hormones between the different immunoassays and in relation to the MS-based reference results (Holst et al., 2004; Wang et al., 2004; Coucke et al., 2007; Soldin and Soldin, 2009; French, 2013; Schofield et al., 2017; Zhou et al., 2017). These discrepancies in immunoassay results indicate differences in the specificities of the antibodies used or inappropriate tracers in the competitive assay formats as well as a lack of standardization of the measurement methods. Efforts to standardize immunoassay methods with respect to MS-based reference methods have been underway for many years (Vesper et al., 2008; Vesper and Botelho, 2010; Vesper and Botelho, 2012; Greaves et al., 2016). Moreover, certified reference materials (CRM) for testosterone, progesterone and 17β-estradiol measurements have been existing for several years (Koumantakis, 2008; Zhou et al., 2017; NIST, updated 2020) and can be used to standardize the respective immunoassays.

This study examines whether these standardization efforts have led to an improvement in testosterone, progesterone and 17β-estradiol immunoassay analytics in recent years. The analysis is based on the manufacturer-specific results of an external quality assessment (EQA) scheme conducted by INSTAND - Society for Promoting Quality in Medical Laboratories e.V. between early 2020 and May 2022.

2 Materials and methods

2.1 Sample materials—preparation and properties

In each EQA survey, two serum samples with different concentrations of testosterone, progesterone and 17β-estradiol were distributed to the participating laboratories for quantitative analysis. The specific analyte concentrations were obtained by spiking pooled human sera with synthetic steroid hormones. The material was stabilized with 0.02% sodium azide and sampled in 2 mL aliquots. The stability and homogeneity of the EQA samples were in line with DIN EN ISO/IEC 17043:2010. The liquid samples were stored at −18°C until they were dispatched to participants at ambient temperature.

2.2 Reference measurement procedure

Reference measurement procedures (RMP) are internationally recognized analytical methods of the highest metrological order, making the reference measurement value (RMV) ideally qualified as a target value for the evaluation of laboratory performances in external quality controls. The RMVs for testosterone, progesterone and 17β-estradiol were determined by the INSTAND calibration laboratory, which is accredited according to DIN EN ISO/IEC 17025:2018 and DIN EN ISO/IEC 15195: 2019. As established RMP for the three steroid hormones, isotope dilution GC-MS (GC-ID/MS) was used. Metrological traceability was established using primary reference standards (Testosterone NMIJ CRM 6002-a, progesterone NMIJ CRM 6003-a, 17β-estradiol NMIJ CRM 6004-a). In order to assign testosterone values, samples were spiked gravimetrically with $^{13}C_2$-testosterone as the internal standard and equilibrated, then precipitated with aqueous KOH, centrifuged, and the supernatant was extracted into dichloromethane. Derivatization was performed with cyclohexane-HFBA and subsequently extracted into cyclohexane phase. GC-MS measurements were done at m/z 680 and m/z 682 (Thienpont et al., 1994). For progesterone measurements, samples were spiked gravimetrically with $^{13}C_2$-progesterone as the internal standard and equilibrated, then extracted into n-hexane. This was followed by centrifugation and evaporation of the supernatant to dryness. Derivatization was performed with HFBA in cyclohexane. GC-MS measurements were done at m/z 510 and m/z 512. In order to assign target values for 17β-estradiol, the samples were spiked gravimetrically with $^{13}C_2$-estradiol as the internal standard, equilibrated, then extracted into dichloromethane, followed by a clean-up step with Sephadex LH-20. Derivatization was performed with cyclohexane/acetone/HFBA. The GC-MS measurements were done at m/z 664 and m/z 666 (Siekmann, 1984). Six measurements were performed for each target value (two measurements per day on three consecutive days). Measurement uncertainty was assigned to each target value on the basis of a measurement uncertainty budget.

2.3 EQA procedure

The INSTAND EQA scheme for measuring testosterone, progesterone and 17β-estradiol is conducted worldwide six times a year (surveys T1 to T6). Two serum samples with two different concentrations (see Section 2.1.) are used per survey (samples S1 and S2). The participating laboratories determine concentrations of testosterone, progesterone, and 17β-estradiol and report on their results via the platform RV-Online (http://rv-online.instandev.de). In addition to submitting the quantitative results for the three steroid hormones, participants are to provide INSTAND with information on the respective device, reagent and method used.

As an RMP is available for testosterone, progesterone and 17β-estradiol, the RMV served as the target value for the evaluation of the EQA results, regardless of the test assays or devices used by the laboratories. For all three steroid hormones, the EQA passing criterion for certification was a deviation from the target value of no more than ±35% according to the rules set out in the guideline of the German Medical Association for quality assurance of medical laboratory analyses (Rili-BÄK) (Bundesärztekammer, 2023).

2.4 Data analysis and statistics

The EQA results for testosterone, progesterone and 17β-estradiol were analyzed for the manufacturer collectives for surveys 2020-T1 to 2022-T3. The number of reported results were generally low for the T2 surveys, making a manufacturer-specific analysis statistically less meaningful. Therefore, only the five other surveys (T1, T3 - T6) were considered in this study. Accordingly, the raw data of twelve surveys in total were analyzed.

Values that scattered farther than 4-fold the standard deviation (SD) of the various collectives were defined as outliers and excluded from the statistical analysis. This definition of outliers primarily excludes gross errors from the analysis that are most likely due to a sample mix-up or a reporting error by individual participants. Thus, ten testosterone results, fourteen progesterone results, and thirteen 17β-estradiol results were excluded (for raw data see Supplementary Table S1).

For all three analytes, the test manufacturer collectives with the highest number of participants per survey were considered, i.e., Abbott (AB), bioMérieux (AX), Siemens and Roche (RO). Siemens consisted of five sub-collectives that showed discrepant results. Therefore, the Bayer Healthcare (SI (BG)) and DPC Biermann (SI (DG)) collectives were presented separately in the analyses. The Dade Behring (SI (BW)), the Siemens Healthineers (SIE) and the Siemens Medical Solutions Diagnostics (SI) collectives had only sporadic participants and were excluded from the analyses. In the case of testosterone, the rather small Tosoh Bioscience (TH) collective was also included as the number of participants increased over the period under observation. See the raw data for details on the assays and devices used by the participating laboratories (Supplementary Table S1).

The distribution of the manufacturer-specific inter-laboratory results for testosterone, progesterone and 17β-estradiol were presented longitudinally as boxplot diagrams. The whiskers of the boxes were defined to stretch from the first quartile −1.5 × (interquartile range) to the third quartile +1.5 × (interquartile range). Further statistical information is provided in Supplementary Table S2. As an RMP is available for all three analytes, the assay-dependent deviations from the RMV were calculated for the EQA results and normalized to the RMV, hereafter designated as bias. The distributions of the bias results for testosterone, progesterone and 17β-estradiol were visualized as boxplot diagrams for sample 2. The normalized manufacturer-dependent biases were examined in relation to the EQA evaluation criterion of ±35% for all three steroid hormones in accordance with the Rili-BÄK guideline (Bundesärztekammer, 2023).

The distribution of the absolute EQA results for the three steroid hormones is provided in the (Supplementary Figure S1).

The EQA results were correlated with the RMV in order to check whether the relative bias of individual manufacturer collectives might indicate a concentration dependency. The manufacturer-specific regression lines could be compared with the y (RMV) = RMV reference line as well as the lower and upper EQA limit of ±35%.

In order to obtain an impression of the value scatter within the individual manufacturer collectives, the coefficients of variation (CV) were calculated for all three steroid hormones.

Basic statistical analyses were performed using JMP 17.0.0 from SAS Institute (Cary, North Carolina, United States).

2.5 Image generation

The overlay images were generated using the Gnu image manipulator software 2.10.8.

3 Results

This study evaluates the quality of inter-laboratory measurements of testosterone, progesterone and 17β-estradiol conducted between early 2020 and May 2022. In a total of twelve EQA surveys, 2,972 results for testosterone, 2,146 for progesterone and 2,292 for 17β-estradiol were reported by 280 participating

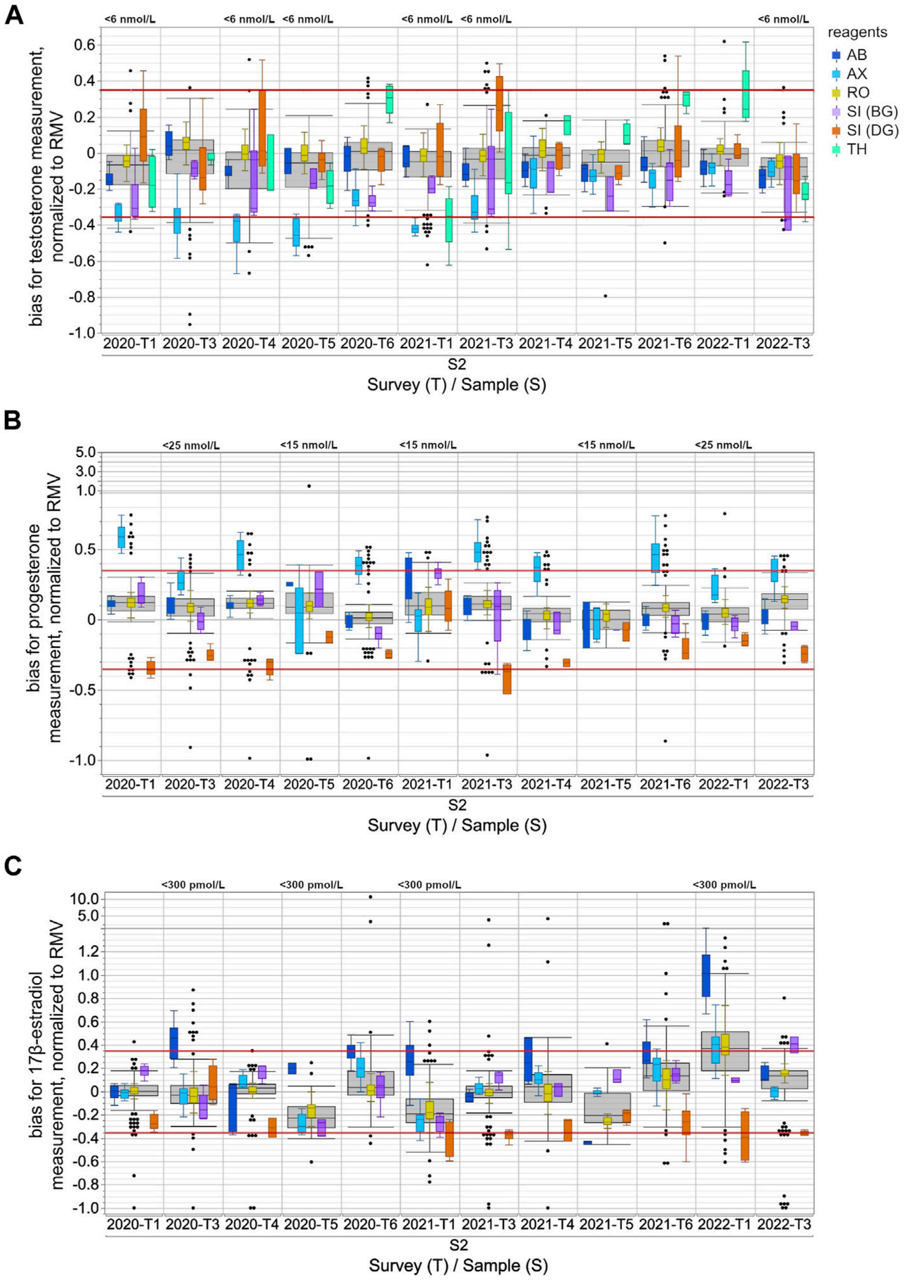


FIGURE 1
Assay-dependent EQA data for testosterone **(A)**, progesterone **(B)** and 17β-estradiol **(C)** measurements in human sera from 2020-T1 to 2022-T3, normalized to the respective reference measurement value (RMV). Only the results for the S2 samples are shown and are representative of all samples. The surveys with EQA samples with low concentrations, testosterone level <6 nmol/L, progesterone level <15 nmol/L or <25 nmol/L, and 17β-estradiol level <300 pmol/L, are labeled in the upper part of the boxplot diagram. Total data is shown as a grey box for the respective survey. The colored boxes show the manufacturer-specific EQA results. The horizontal red line represents the EQA criterion of ±35% of the target value, as determined by reference measurement procedure. For all boxes, the whiskers stretch from the first quartile −1.5 × (interquartile range) to the third quartile +1.5 × (interquartile range). Values outside of this range are shown as dots, but only for the overall results.

laboratories (Supplementary Table S1). After selecting the collectives and eliminating outliers (see Section 2.4.), 2,314 results for testosterone, 1,743 results for progesterone and 1,904 results for 17β-estradiol from 128 laboratories were presented graphically (Supplementary Table S2).

High variation within the manufacturer collectives was found for the three steroid hormones throughout the period analyzed. The whisker ranges reveal that the results of the different collectives do not overlap for some EQA samples (Supplementary Figure S1). While the individual manufacturer collectives showed a clear trend towards increased or decreased levels compared to the overall results for testosterone and progesterone detection, there was a concentration-dependent bias for 17β-estradiol determination (Figure 1C, Supplementary Figure S1C).

When normalizing the results of the individual EQA surveys to the RMV, the overall results for testosterone showed a slight

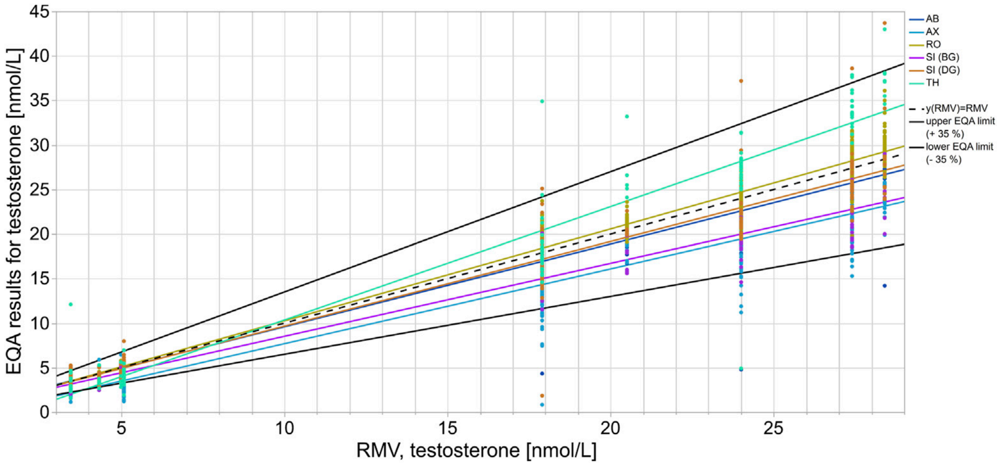


FIGURE 2
Assay-dependent EQA data as represented here by testosterone quantitation correlated to the reference measurement value (RMV). Each color shows the EQA results of a specific assay collective with the respective regression line. The y (RMV) = RMV correlation line is shown as a reference line (black dashes). The solid black lines represent the accepted EQA criterion of ±35%.

tendency towards underestimation, while for progesterone there was a slight tendency towards overestimation (Figure 1). These tendencies seemed to be partly caused by the deviation of the AX collective, which often exceeded the EQA limit of ±35% of the RMV.

In the case of testosterone, the median of the AX collective consistently showed clear deviations from the RMV of −19.7% to −52.2% for all EQA samples up until 2020-T6 (Figure 1A). After 2021-T6, the median of the AX collective deviated less from the RMV for most EQA samples and was even consistently less than −25%. The SI (BG) collectives showed a lower median than the RMV, with a bias down to −36.6% for several EQA surveys. The median bias of the TH collective varied between −35.0% and +32.4%. Interestingly, the upward deviations were only observed in samples with testosterone concentrations above 20 nmol/L (Supplementary Figure S1A). For samples with lower concentrations, the median bias of the TH collective tended to be negative. A correlation of the inter-laboratory test results with the RMV and a comparison of the manufacturer-specific regression lines with the y (RMV) = RMV identity line confirmed that the bias of the TH collective was concentration dependent (Figure 2). A slighter concentration dependency could also be assumed for the AX collective when the regression line was compared with the −35% EQA limit, since a higher percentage deviation was found for low-concentration testosterone samples than for high-concentration ones.

For progesterone, the median bias of the AX collective was often observed to be above the +35% EQA criterion and even up to +58.9% for sample S2 in 2020-T1 (Figure 1B). In individual EQA surveys, the SI (DG) collective median was also slightly below the −35% EQA criterion.

In the case of 17β-estradiol, the overall results showed the highest upward and downward median bias compared to the median bias for testosterone and progesterone measurement (Figure 1C). Upward deviations of the median of the AB collective were observed for 17β-estradiol concentrations below 600 pmol/L, while for higher concentrations, the results were either closer to the RMW or showed a downward deviation. The results of the SI (BG) collective were remarkably high in the case of 17β-estradiol concentrations above 1,000 pmol/L (Supplementary Figure S1C). In contrast, the medians of the SI (DG) collective were consistently low for all EQA samples regardless of the concentration. However, it should be noted that, over the analyzed period, there was a trend towards more negative deviations in the medians of the SI (DG) collective. Since the beginning of 2021, participants of the SI (DG) collective often struggled to meet the −35% EQA criterion (Figure 1C).

For quantitation of all three steroid hormones, the outlier-adjusted CVs were below 25% with a few exceptions for some manufacturer collectives (Figure 3 and Supplementary Figure S3). In the case of testosterone measurement, the CVs were consistently below 10% for the AB and RO collectives. This also applied to the RO collective for progesterone measurement. CVs were consistently below 15% for the AX and RO collectives for 17β-estradiol measurement. Individual cases of remarkably high CVs were observed for various test collectives for all three sex hormones, however these reached a maximum value of 45% (see Supplementary Figure S3B).

4 Discussion

Considering the number of health impairments linked to hormonal disorders (Beneke et al., 2015), reliable and accurate hormone quantitation is essential in order to provide patients with accurate diagnoses and treatment monitoring. However, publications have been reporting for years on the insufficient level of standardization of immunoassays for steroid hormone analysis (Vesper et al., 2008; Vesper and Botelho, 2010; Vesper and Botelho, 2012; Greaves et al., 2016). Certified reference materials are available (Zhou et al., 2017; NIST, updated 2020), but most of the test kit manuals do not provide any information about the traceability of the applied standard samples used to create the

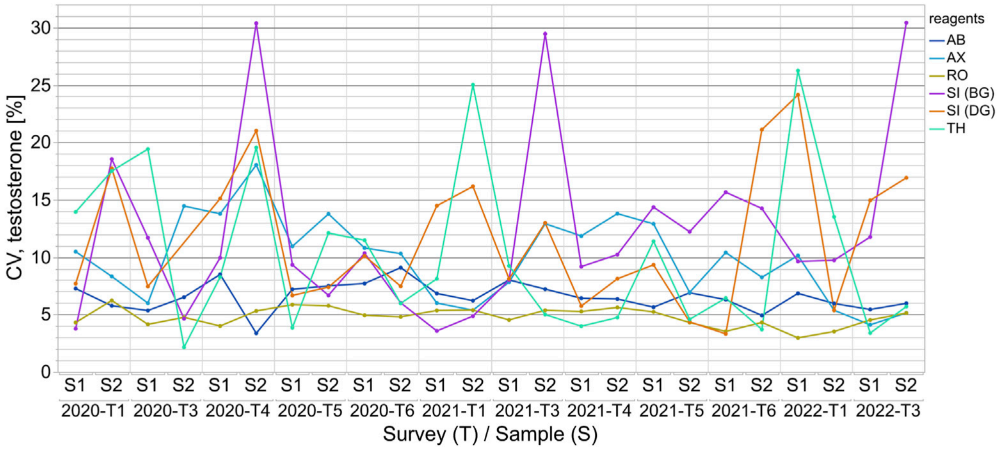


FIGURE 3
The coefficients of variation (CVs) for the assay-dependent EQA results for testosterone measurements from 2020-T1 to 2022-T3 are shown for samples S1 and S2 for each survey. The results of the surveys are independent of one another and thus the CVs are only linked longitudinally to better visualize the changes over time.

respective standard curve or used for 1- or 2-point recalibration. In addition, manufacturers rarely include comparative data with the results of MS-coupled procedures, which are considered the "gold" standard (Krone et al., 2010; Stanczyk and Clarke, 2010; Conklin and Knezevic, 2020). Even though the lack of specificity and selectivity of immunoassays and their disadvantages compared to GC- or LC-MS procedures are well known (Wang et al., 2004; Shackleton, 2010; French, 2013; French, 2016), they are currently still the method of choice in routine measurement as they are practical to carry out and have a high throughput rate. The number of laboratories participating in the EQAs that use MS-coupled methods for steroid hormone determinations has increased in recent years but remains below 10%: around 3% of all 17β-estradiol results, 7% of all testosterone results, and around 8% of all progesterone results (see the raw data in Supplementary Table S1).

This study investigates the quantitative EQA results for testosterone, progesterone, and 17β-estradiol in human serum from twelve INSTAND surveys conducted between early 2020 and May 2022.

The immunoassay-specific results for all three steroid hormones still showed considerable differences. For some EQAs, there was no overlap in the results of different manufacturer collectives when values exceeding the whisker range were disregarded (Figure 1 and Supplementary Figure S1). The EQA results of individual collectives distinctly stood out for progesterone, whereby the overall results of a particular sample overlapped considerably with those of another sample that was twice as concentrated. This was observed with the S2 sample in 2021-T6 and the S1 sample in 2022-T1 (Supplementary Figure S1B).

Normalizing the testosterone, progesterone and 17β-estradiol levels to the RMV allows a comparison to be made of the accuracy of the different immunoassays, even across the several EQA samples and surveys. The median bias of the different collectives was up to approximately 50% for the measurement of both testosterone and 17β-estradiol, and almost 60% for the determination of progesterone (Figure 1). In the case of the 17β-estradiol measurement, the S2 sample in 2022-T1 proved to be an exception with a considerably higher percentage deviation between the manufacturer-dependent results. While both Siemens sub-collectives had similar median biases compared to the other EQA samples, the other three collectives showed substantially higher upward deviations. One can assume that a cross-reacting compound in this particular sample interfered with the measurement of 17β-estradiol in the AX, RO, and especially the AB immunoassays (Sturgeon and Viljoen, 2011; Wauthier et al., 2022), however the compound did not interfere with the measurement of testosterone or progesterone (Figure 1). An interfering substance in an EQA sample may be either of endogenous origin in the serum matrix or due to artificial additives which are used during sample preparation for the purpose of stabilization or spiking. Since the manufacturing process of the EQA sample remained the same for all of the analyzed EQA surveys, it can be assumed this was not caused by an artificial additive in this sample. The fact that test kits from other manufacturers were not impacted by this presumably interfering compound shows that the immunoassays in these kits may be more effectively protected against cross-reacting substances than the methods mentioned above.

The high structural and steric similarity of the numerous derivates in the steroid family means that differentiation by antibody detection is difficult due to cross-reactivity (Krasowski et al., 2014; Yamamoto et al., 2014; Beneke et al., 2015) and thus poses a major challenge for the immunoassay measurement of steroid hormones. Test manufacturers list several cross-reacting molecules in their test manuals, e.g., in progesterone analyses, the rate of a cross-reaction with 11-deoxycorticosterone is 1%–4% depending on the test. In the test manuals for testosterone measurement, much higher interference rates of up to 34% are reported for 11β-hydroxy-testosterone and 11-keto-testosterone. Krasowski et al. found higher cross-reactivities for testosterone measurement than for progesterone and 17β-estradiol determination in the Roche Diagnostics Elecsys assays (Krasowski et al., 2014).

The many possible interfering substances can lead to both an over- and underestimation of testosterone, progesterone and 17βestradiol levels (Sturgeon and Viljoen, 2011). In general, overestimated steroid hormone levels in serum can result in the erroneous diagnosis of hormonal diseases and cause avoidable uncertainty among patients. Underestimated sex hormone levels can falsely lead to a presumed case of hypogonadism and, in turn, unnecessary hormone substitution in patients (Zitzmann and Nieschlag, 2000; Zitzmann et al., 2006). To avoid misdiagnoses, hormone measurements should be interpreted with caution, especially for patients on medication, since cross-reactivity occurs with drugs that have a high structural similarity, e.g., with methyltestosterone in some testosterone immunoassays (Krasowski et al., 2014).

As a consequence, the same immunoassay should be used for patient monitoring and follow-up in order to minimize discrepant results and uncertainty for clinicians and patients due to possible assay-dependent under- or overdetermination in steroid hormone measurement.

Most manufacturer collectives deviated either upwards or downwards from the RMV when quantifying steroid hormones, however some collectives showed deviations from the RMV in both directions (Figure 1). In the case of 17β-estradiol quantitation, positive as well as negative biases to the RMV were observed for all manufacturer collectives, as well as for the total collective. In these cases, the deviations of the assay collective seemed to depend on the hormone concentration in the EQA sample (Figure 2, Supplementary Figures S1, S2).

The testosterone results for the TH collective were remarkably higher than the RMV for samples with high concentrations, e.g., sample S2 in 2020-T6, 2021-T4, 2021-T6 and 2022-T1. In contrast, samples with concentrations below 6 nmol/L were underestimated, see sample S2 in 2020-T1, 2020-T4, 2020-T5, 2021-T1, 2021-T3 and 2022-T3 (Figure 1 and Supplementary Figure S1). This concentration dependency might be due to an imprecise test calibration or due to insufficient sensitivity in cases of low steroid hormone concentrations. However, the testosterone concentrations of the EQA samples were within the measuring ranges specified in the test manuals of the assay manufacturers and were within clinically relevant concentrations (Beneke et al., 2015).

Kanakis and others found that most commercially available immunoassays used for testosterone quantitation are insufficient for lower concentrations within the normal reference range for men (~10 nmol/L to ~35 nmol/L) and the entire reference range for women (~0.2 nmol/L to ~3 nmol/L). For this reason, slight androgen excess in female patients cannot be measured by some of the commercial tests and remains undetected (Kanakis et al., 2019). This problem is addressed, for example, in EQA samples S1 in 2020-T3 and S2 in 2021-T3 representing elevated female serum testosterone levels. These elevated levels would likely not be identified using the AX or the TH immunoassays due to underestimation (Supplementary Figure S1A). This can result in an unreliable diagnosis of diseases associated with androgen excess in women, such as idiopathic hirsutism, PCOS, hyperthecosis ovarii, late-onset congenital adrenal hyperplasia or testosterone-producing tumors. Some groups reported challenges in measuring low serum testosterone concentrations (La'ulu et al., 2018; Cao et al., 2019; Kanakis et al., 2019). La'ulu et al. described sensitivities for testosterone measurement with various commercial immunoassays in concentrations ranging from 0.36 nmol/L to 3.49 nmol/L (Schwartz et al., 1986; Legro et al., 2013; Beneke et al., 2015; Azziz, 2018; Cussen et al., 2022). On the other hand, samples with low levels of steroid hormones can also be overestimated, as interfering substances and cross-reactivities could overwhelm the measurement of the target analyte. This would result in unrecognized hypogonadism in patients (Corona et al., 2011; Skałba and Guz, 2011; Basaria, 2014; Beneke et al., 2015; Klein et al., 2017).

The same challenges arise when measuring low concentrations of progesterone (<5 nmol/L) and 17β-estradiol (<40.7 pmol/L) (Oettel and Mukhopadhyay, 2004; Huhtaniemi et al., 2012; Shankara-Narayana et al., 2016) in male patients or in women with depressed levels. For EQA result distribution for EQA samples with progesterone concentrations <5 nmol/L see also sample S2 in 2020-T5 and 2021-T2 (Figure 1). For 17β-estradiol, the lowest concentrations in the EQA scheme were around 150 pmol/L, e.g., sample S2 in 2021-T1 and sample S1 in 2021-T5. The EQA results reveal clear measurement differences between the individual collectives (Figure 1 and Supplementary Figure S1). All in all, an improvement in immunoassay measurements is especially desirable for samples with low hormone levels and should be pursued further by the current standardization programs.

While the wide variations within the manufacturer collectives in testosterone, progesterone and 17β-estradiol immunoassay measurement revealed issues with accuracy, within-assay agreement was mainly good, indicating relatively good analytical precision. The outlier-adjusted scatter within the collectives was found to be mostly inconspicuous (Figure 3, Supplementary Figure S3) and similar to the manufacturer's specifications in the test manuals. The CVs for the manufacturer collectives were, with few exceptions, below 15% for all three steroid hormones. For testosterone quantitation, slightly higher CVs were observed for the SI (BG), SI (DG) and TH collectives than for the others. This was most likely due to the lower number of EQA results for these collectives. In the case of progesterone and 17β-estradiol determination, the CVs for SI (DG) and SI (BG), and, in the case of 17β-estradiol, for the AB collective as well, should be interpreted with caution for the same reason. For all three hormones and all test collectives, slightly increased CVs were observed over two to three consecutive surveys. One possible explanation for this could be lot changes by manufacturers.

Overall, the bias analysis of the testosterone, progesterone and 17β-estradiol data confirmed the findings of previously published studies, which found that immunoassays were insufficiently reliable in quantitatively determining sex hormones. A trend towards standardizing immunoassay detection for steroids has yet to be observed (Vesper et al., 2014; Lawrenz et al., 2018). However, this EQA data revealed one exception. The dispersion of testosterone values between the different assays decreased over the studied period. This can be ascribed to the development towards a higher accuracy in the AX collective. Until 2021-T3, the results of the AX collective had often exceeded the EQA criterion of −35%. Since the beginning of 2021, the median of the AX collective has remarkably moved closer to the RMV (Figure 1A). This improvement in accuracy could be due to a successful recalibration by the manufacturer. Test system recalibrations have to be performed under consideration of traceability (Koumantakis, 2008). The increased accuracy in testosterone quantitation for the AX immunoassay since 2021 is a good example of how external quality control schemes can reveal inadequate test performance, a matter which can subsequently be discussed with the manufacturers. This can ultimately help improve analytics and thus promote quality assurance in medical laboratories.

One limitation of this study is that stabilized and spiked serum samples were used for the EQAs. However, since manufacturer-dependent deviations in steroid hormone measurements are also described for fresh serum samples in other studies (Taieb et al., 2003; Wang et al., 2004; Coucke et al., 2007; Bell et al., 2012; Cao et al., 2019), it is rather unlikely that the observed manufacturer-specific deviations in the EQA results are primarily due to insufficient commutability of the EQA samples. To make sure that the manufacturer-dependent deviations from the RMV were not, or only negligibly, influenced by the artificial nature of the samples, INSTAND will address this aspect in detail in further studies by providing fresh, non-processed serum samples.

5 Conclusion

While the scatter within the manufacturer collectives of the EQA was not critical for the quantitation of testosterone, progesterone and 17β-estradiol using immunoassays, there were considerable differences between the manufacturer-specific EQA results. This revealed the need for distinct improvement in standardization. The increased accuracy of the AX immunoassay in measuring testosterone in the INSTAND EQAs between 2020 and 2022 might be due to successful recalibration of the assay and raises hope for further improvement in the standardization of immunoassays for steroid hormone analysis in the coming years.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

LV: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing–original draft. NW: Writing–review and editing. PK: Supervision, Writing–review and editing. MK: Writing–review and editing. FW: Writing–review and editing. MT: Writing–review and editing. IS: Project administration, Resources, Supervision, Writing–review and editing. PL: Supervision, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2024.1345356/full#supplementary-material

References

Armeni, E., Paschou, S. A., Goulis, D. G., and Lambrinoudaki, I. (2021). Hormone therapy regimens for managing the menopause and premature ovarian insufficiency. *Best Pract. Res. Clin. Endocrinol. Metabolism* 35 (6), 101561. doi:10.1016/j.beem.2021.101561

Aubard, Y., Teissier, M. P., Grandjean, M. H., Le Meur, Y., and Baudet, J. H. (1997). Early menopause. *J. Gynecol. Obstet. Biol. Reprod. Paris.* 26 (3), 231–237.

Azziz, R. (2018). Polycystic ovary syndrome. *Obstetrics Gynecol.* 132 (2), 321–336. doi:10.1097/aog.0000000000002698

Barbonetti, A., D'Andrea, S., and Francavilla, S. (2020). Testosterone replacement therapy. *Andrology* 8 (6), 1551–1566. doi:10.1111/andr.12774

Basaria, S. (2014). Male hypogonadism. *Lancet* 383 (9924), 1250–1263. doi:10.1016/S0140-6736(13)61126-5

Bell, A., Meek, C. L., and Viljoen, A. (2012). Evidence of biochemical hyperandrogenism in women: the limitations of serum testosterone quantitation. *J. Obstetrics Gynaecol.* 32 (4), 367–371. doi:10.3109/01443615.2012.656741

Beneke, H., Claudi-Böhm, S., Gerhardt, A., Manfras, B., and Plonné, D. (2015). *Klinikleitfaden labordiagnostik*.

Bundesärztekammer (2023). Richtlinie der Bundesärztekammer zur Qualitätssicherung laboratoriumsmedizinischer Untersuchungen. Available at: https://www.bundesaerztekammer.de/fileadmin/user_upload/BAEK/Themen/Qualitaetssicherung/_Bek_BAEK_RiLi_BAEK_ONLINE_FINAL_VERS_26_05_2023.pdf (Accessed August 16, 2023).

Cao, Z., Botelho, J. C., Rej, R., Vesper, H., and Astles, J. R. (2019). Impact of testosterone assay standardization efforts assessed via accuracy-based proficiency testing. *Clin. Biochem.* 68, 37–43. doi:10.1016/j.clinbiochem.2019.03.014

Conklin, S. E., and Knezevic, C. E. (2020). Advancements in the gold standard: measuring steroid sex hormones by mass spectrometry. *Clin. Biochem.* 82, 21–32. doi:10.1016/j.clinbiochem.2020.03.008

Corona, G., Rastrelli, G., Morelli, A., Vignozzi, L., Mannucci, E., and Maggi, M. (2011). Hypogonadism and metabolic syndrome. *J. Endocrinol. Investigation* 34 (7), 557–567. doi:10.3275/7806

Coucke, W., Devleeschouwer, N., Libeer, J. C., Schiettecatte, J., Martin, M., and Smitz, J. (2007). Accuracy and reproducibility of automated estradiol-17beta and progesterone assays using native serum samples: results obtained in the Belgian external assessment scheme. *Hum. Reprod.* 22 (12), 3204–3209. doi:10.1093/humrep/dem322

Cussen, L., McDonnell, T., Bennett, G., Thompson, C. J., Sherlock, M., and O'Reilly, M. W. (2022). Approach to androgen excess in women: clinical and biochemical insights. *Clin. Endocrinol. (Oxf)* 97 (2), 174–186. doi:10.1111/cen.14710

Diemer, T., Hauptmann, A., and Wagenlehner, F. M. (2016). Testosterone therapy. *Urol. A* 55 (4), 539–548. doi:10.1007/s00120-016-0072-y

French, D. (2013). Development and validation of a serum total testosterone liquid chromatography-tandem mass spectrometry (LC-MS/MS) assay calibrated to NIST SRM 971. *Clin. Chim. Acta* 415, 109–117. doi:10.1016/j.cca.2012.10.007

French, D. (2016). Advances in bioanalytical techniques to measure steroid hormones in serum. *Bioanalysis* 8 (11), 1203–1219. doi:10.4155/bio-2015-0025

Gleicher, N., Brown, T., Dudkiewicz, A., Karande, V., Rao, R., Balin, M., et al. (2000). Estradiol/progesterone substitution in the luteal phase improves pregnancy rates in stimulated cycles--but only in younger women. *Early Pregnancy (Cherry Hill)* 4 (1), 64–73.

Greaves, R. F., Ho, C. S., Hoad, K. E., Joseph, J., McWhinney, B., Gill, J. P., et al. (2016). Achievements and future directions of the APFCB mass spectrometry harmonisation project on serum testosterone. *Clin. Biochem. Rev.* 37 (2), 63–84.

Holst, J. P., Soldin, O. P., Guo, T., and Soldin, S. J. (2004). Steroid hormones: relevance and measurement in the clinical laboratory. *Clin. Lab. Med.* 24 (1), 105–118. doi:10.1016/j.cll.2004.01.004

Huhtaniemi, I. T., Tajar, A., Lee, D. M., O'Neill, T. W., Finn, J. D., Bartfai, G., et al. (2012). Comparison of serum testosterone and estradiol measurements in 3174 European men using platform immunoassay and mass spectrometry; relevance for the diagnostics in aging men. *Eur. J. Endocrinol.* 166 (6), 983–991. doi:10.1530/eje-11-1051

Kanakis, G. A., Tsametis, C. P., and Goulis, D. G. (2019). Measuring testosterone in women and men. *Maturitas* 125, 41–44. doi:10.1016/j.maturitas.2019.04.203

Klein, D. A., Emerick, J. E., Sylvester, J. E., and Vogt, K. S. (2017). Disorders of puberty: an approach to diagnosis and management. *Am. Fam. Physician* 96 (9), 590–599.

Kleine, B., and Rossmanith, W. (2013). *Hormone und Hormonsystem - lehrbuch der Endokrinologie*. Heidelberg: Springer Spektrum Berlin.

Koumantakis, G. (2008). Traceability of measurement results. *Clin. Biochem. Rev.* 29 (Suppl. 1), S61–S66.

Krasowski, M. D., Drees, D., Morris, C. S., Maakestad, J., Blau, J. L., and Ekins, S. (2014). Cross-reactivity of steroid hormone immunoassays: clinical significance and two-dimensional molecular similarity prediction. *BMC Clin. Pathol.* 14 (33), 33. doi:10.1186/1472-6890-14-33

Krone, N., Hughes, B. A., Lavery, G. G., Stewart, P. M., Arlt, W., and Shackleton, C. H. (2010). Gas chromatography/mass spectrometry (GC/MS) remains a pre-eminent discovery tool in clinical steroid investigations even in the era of fast liquid chromatography tandem mass spectrometry (LC/MS/MS). *J. Steroid Biochem. Mol. Biol.* 121 (3-5), 496–504. doi:10.1016/j.jsbmb.2010.04.010

La'ulu, S. L., Kalp, K. J., and Straseski, J. A. (2018). How low can you go? Analytical performance of five automated testosterone immunoassays. *Clin. Biochem.* 58, 64–71. doi:10.1016/j.clinbiochem.2018.05.008

Lawrenz, B., Sibal, J., Garrido, N., Abu, E., Jean, A., Melado, L., et al. (2018). Inter-assay variation and reproducibility of progesterone measurements during ovarian stimulation for IVF. *PLoS One* 13 (11), e0206098. doi:10.1371/journal.pone.0206098

Legro, R. S., Arslanian, S. A., Ehrmann, D. A., Hoeger, K. M., Murad, M. H., Pasquali, R., et al. (2013). Diagnosis and treatment of polycystic ovary syndrome: an endocrine society clinical practice guideline. *J. Clin. Endocrinol. Metabolism* 98 (12), 4565–4592. doi:10.1210/jc.2013-2350

NIST updated (2020). Development of reference methods and reference materials for the determination of hormones in human serum. Available at: https://www.nist.gov/programs-projects/development-reference-methods-and-reference-materials-determination-hormones-human (Accessed August 1, 2023).

Oettel, M., and Mukhopadhyay, A. K. (2004). Progesterone: the forgotten hormone in men? *Aging Male* 7 (3), 236–257. doi:10.1080/13685530400004199

Salonia, A., Rastrelli, G., Hackett, G., Seminara, S. B., Huhtaniemi, I. T., Rey, R. A., et al. (2019). Paediatric and adult-onset male hypogonadism. *Nat. Rev. Dis. Prim.* 5 (1), 38. doi:10.1038/s41572-019-0087-y

Schofield, R. C., Mendu, D. R., Ramanathan, L. V., Pessin, M. S., and Carlow, D. C. (2017). Sensitive simultaneous quantitation of testosterone and estradiol in serum by LC-MS/MS without derivatization and comparison with the CDC HoSt program. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* 1048, 70–76. doi:10.1016/j.jchromb.2017.02.006

Schwartz, U., Moltz, L., Pickartz, H., Sörensen, R., and Römmler, A. (1986). Hyperthecosis ovarii--a tumor-like change in androgenized females. *Geburtshilfe Frauenheilkd* 46 (06), 391–397. doi:10.1055/s-2008-1035937

Shackleton, C. (2010). Clinical steroid mass spectrometry: a 45-year history culminating in HPLC-MS/MS becoming an essential tool for patient diagnosis. *J. Steroid Biochem. Mol. Biol.* 121 (3-5), 481–490. doi:10.1016/j.jsbmb.2010.02.017

Shankara-Narayana, N., Zawada, S., Walters, K. A., Desai, R., Marren, A., and Handelsman, D. J. (2016). Accuracy of a direct progesterone immunoassay. *J. Appl. Lab. Med.* 1 (3), 294–299. doi:10.1373/jalm.2016.020123

Siekmann, L. (1984). Determination of oestradiol-17ß in human serum by isotope dilution-mass spectrometry. Definitive methods in clinical chemistry, II. *Defin. Methods Clin. Chem. II.* 22(8), 551–558. doi:10.1515/cclm.1984.22.8.551

Skałba, P., and Guz, M. (2011). Hypogonadotropic hypogonadism in women. *Endokrynol. Pol.* 62 (6), 560–567.

Soldin, S. J., and Soldin, O. P. (2009). Steroid hormone analysis by tandem mass spectrometry. *Clin. Chem.* 55 (6), 1061–1066. doi:10.1373/clinchem.2007.100008

Stanczyk, F. Z., and Clarke, N. J. (2010). Advantages and challenges of mass spectrometry assays for steroid hormones. *J. Steroid Biochem. Mol. Biol.* 121 (3-5), 491–495. doi:10.1016/j.jsbmb.2010.05.001

Strawn, E. Y., Roesler, M., Rinke, M., and Aiman, E. J. (2000). Minimal precycle testing and ongoing cycle monitoring for *in vitro* fertilization and fresh pre-embryo transfer do not compromise fertilization, implantation, or ongoing pregnancy rates. *Am. J. Obstet. Gynecol.* 182 (6), 1623–1628. doi:10.1067/mob.2000.107434

Sturgeon, C. M., and Viljoen, A. (2011). Analytical error and interference in immunoassay: minimizing risk. *Ann. Clin. Biochem.* 48 (5), 418–432. doi:10.1258/acb.2011.011073

Taieb, J., Mathian, B., Millot, F., Patricot, M. C., Mathieu, E., Queyrel, N., et al. (2003). Testosterone measured by 10 immunoassays and by isotope-dilution gas chromatography-mass spectrometry in sera from 116 men, women, and children. *Clin. Chem.* 49 (8), 1381–1395. doi:10.1373/49.8.1381

Thienpont, L. M., De Brabandere, V. I., Stöckl, D., and De Leenheer, A. P. (1994). Use of cyclodextrins for prepurification of progesterone and testosterone from human serum prior to determination with isotope dilution gas chromatography/mass spectrometry. *Anal. Chem.* 66 (22), 4116–4119. doi:10.1021/ac00094a041

Thomsen, L. H., Humaidan, P., Erb, K., Overgaard, M., Andersen, C. Y., and Kesmodel, U. S. (2018). Mid-Luteal 17-OH progesterone levels in 614 women undergoing IVF-treatment and fresh embryo transfer-daytime variation and impact on live birth rates. *Front. Endocrinol. (Lausanne)* 9, 690. doi:10.3389/fendo.2018.00690

Vesper, H. W., and Botelho, J. C. (2010). Standardization of testosterone measurements in humans. *J. Steroid Biochem. Mol. Biol.* 121 (3), 513–519. doi:10.1016/j.jsbmb.2010.03.032

Vesper, H. W., and Botelho, J. C. (2012). Testosterone. An overview of CDC's standardization initiative. *Clin. Lab. News.*

Vesper, H. W., Botelho, J. C., Shacklady, C., Smith, A., and Myers, G. L. (2008). CDC project on standardizing steroid hormone measurements. *Steroids* 73 (13), 1286–1292. doi:10.1016/j.steroids.2008.09.008

Vesper, H. W., Botelho, J. C., and Wang, Y. (2014). Challenges and improvements in testosterone and estradiol testing. *Asian J. Androl.* 16 (2), 178–184. doi:10.4103/1008-682X.122338

Wang, C., Catlin, D. H., Demers, L. M., Starcevic, B., and Swerdloff, R. S. (2004). Measurement of total serum testosterone in adult men: comparison of current laboratory methods versus liquid chromatography-tandem mass spectrometry. *J. Clin. Endocrinol. Metab.* 89 (2), 534–543. doi:10.1210/jc.2003-031287

Wauthier, L., Plebani, M., and Favresse, J. (2022). Interferences in immunoassays: review and practical algorithm. *Clin. Chem. Laboratory Med. (CCLM)* 60 (6), 808–820. doi:10.1515/cclm-2021-1288

Yadav, V., and Sharma, Y. (2023). Hyperandrogenism. *Indian J. Pediatr.* 90 (10), 1018–1024. doi:10.1007/s12098-023-04678-7

Yamamoto, K., Kohama, M., Nakahara, F., Yamakami, A., Tanaka, C., Momoeda, M., et al. (2014). Cross-reactivity evaluation of improved estradiol (E2) assay reagent based on chemiluminescent enzyme immunoassay. *Rinsho Byori* 62 (8), 755–760.

Zhou, H., Wang, Y., Gatcombe, M., Farris, J., Botelho, J. C., Caudill, S. P., et al. (2017). Simultaneous measurement of total estradiol and testosterone in human serum by isotope dilution liquid chromatography tandem mass spectrometry. *Anal. Bioanal. Chem.* 409 (25), 5943–5954. doi:10.1007/s00216-017-0529-x

Zitzmann, M., Faber, S., and Nieschlag, E. (2006). Association of specific symptoms and metabolic risks with serum testosterone in older men. *J. Clin. Endocrinol. Metabolism* 91 (11), 4335–4343. doi:10.1210/jc.2006-0401

Zitzmann, M., and Nieschlag, E. (2000). Hormone substitution in male hypogonadism. *Mol. Cell. Endocrinol.* 161 (1), 73–88. doi:10.1016/S0303-7207(99)00227-0

TYPE Original Research
PUBLISHED 20 March 2024
DOI 10.3389/fmolb.2024.1371426

Check for updates

OPEN ACCESS

EDITED BY
Hem Chandra Jha,
Indian Institute of Technology Indore, India

REVIEWED BY
Scott Pardo,
Diabetes Care Center, United States
Brian Aufderheid,
Hampton University, United States

*CORRESPONDENCE
Peter B. Luppa,
✉ p.luppa@tum.de

RECEIVED 16 January 2024
ACCEPTED 05 March 2024
PUBLISHED 20 March 2024

CITATION
Luppa PB, Zeller M, Pieper M, Kaiser P, Weiss N, Vierbaum L and Freckmann G (2024), Quality assessment of glucose measurement with regard to epidemiology and clinical management of diabetes mellitus in Germany. *Front. Mol. Biosci.* 11:1371426.
doi: 10.3389/fmolb.2024.1371426

Quality assessment of glucose measurement with regard to epidemiology and clinical management of diabetes mellitus in Germany

Peter B. Luppa[1]*, Michael Zeller[1], Marija Pieper[1], Patricia Kaiser[2], Nathalie Weiss[2], Laura Vierbaum[2] and Guido Freckmann[3]

[1]Institut für Klinische Chemie und Pathobiochemie, Klinikum rechts der Isar der Technische Universität München, Munich, Germany, [2]INSTAND e.V., Gesellschaft zur Förderung der Qualitätssicherung in Medizinischen Laboratorien e.V., Düsseldorf, Germany, [3]Institut für Diabetes-Technologie, Forschungs- und Entwicklungsgesellschaft mbH an der Universität Ulm, Ulm, Germany

Background: During the last decade, Germany has seen an increased prevalence and a redistribution from undetected to diagnosed diabetes mellitus. Due to this substantial epidemiological development, the number of people with documented type 2 diabetes was 8.7 million in 2022. An estimated two million undiagnosed subjects are to be added. Beyond that, the life expectancy of diabetic subjects is increasing due to more responsive health systems in terms of care. Possible reasons include improved screening of at-risk individuals, the introduction of HbA1c for diagnosis in 2010, and the higher use of risk scores. Additionally, quality aspects of the laboratory methodology should be taken into consideration.

Methods: Epidemiology and clinical management of diabetes in Germany are presented in the light of publications retrieved by a selective search of the PubMed database. Additionally, the data from German external quality assessment (EQA) surveys for the measurands glucose in plasma and HbA1c in whole blood, reviewed from 2010 until 2022, were evaluated. Above this, data concerning the analytical performance of near-patient glucometer devices, according to the ISO norm 15197:2013, were analyzed.

Results: Two laboratory aspects are in good accordance with the observation of an increase in the diabetes mellitus prevalence when retrospectively reviewing the period 2010 to 2022: First, the analytical performance according to the ISO norm 15197:2013 of the glucometer devices widely used by patients with diabetes for the glucose self-testing, has improved during this period. Secondly, concerning the EQA program of INSTAND, the number of participating laboratories raised significantly in Germany. The spreads of variations of the specified results for plasma glucose remained unchanged between 2010 and 2022, whereas for HbA1c a significant decrease of the result scattering could be observed.

Conclusion: These retrospectively established findings testify to an excellent analytical quality of laboratory diagnostics for glucose and HbA1c throughout Germany which may be involved in a better diagnosis and therapy of previously undetected diabetes mellitus.

KEYWORDS

EQA schemes, diabetes prevalence, diabetes management, ISO 15197, glucose in plasma, HbA1c in whole blood

Introduction

Diabetes mellitus is a group of common endocrine diseases characterized by sustained high plasma glucose and elevated whole blood glycated hemoglobin (HbA1c) concentrations and resulting clinical signs of persistent hyperglycemia. The chronic and untreated life-threatening disease is due to either pancreatic lesions resulting in impaired insulin secretion or peripheral cells becoming unresponsive to insulin to a variable degree and its subsequent metabolic effects (so-called peripheral insulin resistance) (Brutsaert, 2023). The vast majority of affected patients suffer from type 1 and type 2 diabetes. The high worldwide burden of diabetes has adverse health effects on affected individuals, but also economic impacts on the global healthcare systems.

In Germany, seven million people had documented type-2 diabetes in 2015. In the same year, 32,000 children and adolescents, as well as 340,000 adults, had type 1 diabetes. Due to the increasing prevalence data, the number of people with documented type 2 diabetes was expected to reach 8.7 million in 2022 (Tönnies et al., 2019).

Worldwide, diabetes mellitus and the healthcare resources required to treat the disease result in challenging high socio-economic costs. With approx. forty billion €, Germany has the fourth highest healthcare expenditure on diabetes. Healthcare costs for affected patients are around twice as high as for comparable people without diabetes. A large proportion of healthcare expenditure is spent on treating secondary diseases of diabetes. Sophisticated disease management programs can limit the increase in this expenditure (Deutsche Diabetes Gesellschaft, 2023).

Laboratory medical examinations are of great importance in the diagnosis and subsequent disease management of diabetes mellitus (Schleicher et al., 2022). Blood or plasma glucose measurement has long been a proven analytical method performed in the central laboratory, but also near-patient blood glucose measuring devices, which, if subjected to close-controlled quality assurance measures, allow highly accurate determinations of plasma glucose. Furthermore, in the last decade, HbA1c has emerged as a long-term diagnostic parameter in addition to the already-known assessment for glycemic control in people with diabetes. It may complement the determination of glucose in a diagnostically helpful way. The essential role of HbA1c is that it can be used to make a statement about the blood glucose control of the last 8–12 weeks and can thus be applied as a therapy control to reduce possible consequential damage (Weykamp, 2013). In addition to diagnosing new-onset diabetes mellitus, the lifelong monitoring of glucose metabolism is another vital pillar of treating this disease. Most patients carry out this measurement by themselves using glucometers with unit-use test strips daily. This so-called self-measurement of blood glucose (SMBG) has recently been supplemented by continuous glucose monitoring (CGM) systems (Lin et al., 2021), which involves continuous measurement of glucose in the interstitial body fluid. In addition, CGM-controlled and partially automated insulin dosing systems are already on the rise.

Ongoing improvements in SMBG/CGM analytics, insulin injection technology, and data management have evolved into a novel modern form of diabetes treatment alongside education/counseling and adequate drug therapy. Physicians in diabetology-focus practices, outpatient clinics, and hospitals are thus provided with more and more data to assess and optimize the quality of the individual patient's diabetes situation (Kravarusic and Aleppo, 2020).

The accuracy and precision of laboratory parameters undoubtedly have a direct impact on diagnosis and patient care. All measurements, including those of the pivotal parameter plasma glucose concentration, are subject to an inherent measurement uncertainty (Petersmann et al., 2022). Analytical efforts should, therefore, always aim to reduce the measurement uncertainty in order to meet the requirements for the diagnosis and treatment of diabetes. Such efforts have been observed in recent years, including for HbA1c. For this parameter, the permissible relative root mean square measurement error in EQA schemes was reduced from ±10% to ±3% (Bundesärztekammer, 2023). This should improve the analytical differentiation between the important HbA1c cutoff values of 39 and 48 mmol/mol (5.7% and 6.5%).

The aim of this article is, therefore, to highlight the changes in the epidemiology and clinical management of diabetes mellitus that have been achieved over the last 2 decades and to causally relate them to the advances in analytical capabilities and improved quality assurance measures, here, in particular, the External Quality Assessment (EQA) schemes for the measurands glucose and HbA1c, executed in Germany.

The International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) defines EQA as laboratory performance and method evaluation for regulatory purposes, focusing on participating laboratories' or physicians' education and support (Blasutig et al., 2023). EQA primarily evaluates the analytical performance of participants concerning measurands by comparison to a target consensus value (CV) within a method split or by comparison of all methods to a reference method value (RMV). In Germany, EQA schemes are classified as regulatory based on the valid at the time "Guideline of the German Medical Association for the Quality Assurance of Laboratory Medical Examinations (Rili-BÄK)" (Bundesärztekammer, 2023). The Society for Promoting Quality Assurance in Medical Laboratories (INSTAND), Düsseldorf, and the Reference Institute for Bioanalytics (RfB), Bonn, are accredited organizations performing EQA schemes in laboratory medicine.

Strict quality standards for IVD manufacturers are also mandatory for a premarket evaluation to ensure the measurement quality of the respective device. The international standard ISO 15197 (see below) is a norm applied to blood glucose monitoring systems.

Materials and methods

Epidemiological data for diabetes mellitus, gestational diabetes

Epidemiology and clinical management of diabetes in Germany are presented in the light of publications retrieved by a selective

TABLE 1 Characteristics of INSTAND's EQA schemes. RMV indicates the use of a respective reference method value for evaluation, whereas CV stands for consensus value mode.

EQA scheme measurand, additional description	Code #	Average number of participants per survey[a]	EQA evaluation mode	Number of evaluated device types/methods
Glucose as part of the clinical chemistry EQA	100	657	RMV	8/2
Glucose POCT	800	706	CV	28/2
HbA1c	145	701	RMV	17/3 Proportions of methods: 61% immunological; 25% HPLC; 5% affinity chromatography

[a]Per year, there are six EQA surveys for each scheme.

search of the PubMed database (search terms were *diabetes mellitus type 1 AND diabetes mellitus type 2 AND epidemiology AND mortality AND clinical management*), as well as by the annual healthcare reports of the German Diabetes Society (Deutsche Diabetes Gesellschaft, DDG) and its pertinent guidelines (including gestational diabetes).

EQA data for glucose and HbA1c

The second data source was the EQA surveys for the measurands glucose and HbA1c, conducted by INSTAND, from 2010 to 2022. Each survey (synonym "ring trial") is offered six times per year. For the statistical analysis of the quantitative results for glucose and HbA1c, only the last EQA scheme of the year, conducted in October, was used since it was regularly the largest one regarding the number of participants. Concerning our analysis, the EQA schemes discussed are evaluated as follows: EQA #100 (clinical chemistry parameters, including glucose): RMV; #145 (HbA1c): RMV; #800 (glucose POCT): CV. In each EQA survey, two samples with different concentration levels (randomly assigned as samples A and B) are delivered to the participants. These concentrations are chosen to be within the linear measurement range of all possible methods used by the participants.

The total numbers per year were analyzed to determine the dynamics of the participation. We also used the respective RfB information for the corresponding ring trials. In general, all participants must report quantitative results together with additional information on the test kit provider and laboratory equipment used via the online portal of the respective reference institution. Table 1 summarizes the number of participants and applied methods for the INSTAND EQA schemes.

INSTAND offers EQA schemes for glucose in plasma as part of the clinical chemistry panel (#100) or separate as POCT glucose samples (#800). We analyzed the value scattering of the respective results of the EQA #100 (glucose oxidase (GOD) and hexokinase/Glc-6P-DH method) and #800 (GOD and Glc-DH method) throughout the whole period 2010–2022 without differentiation of the methods applied. For HbA1c as the second diabetes measurement, the EQA organization has used fresh whole blood samples since 2015, with target values assigned with the IFCC reference measurement procedure. Therefore, we analyzed this ring trial only in 2015–2022. Here, we differentiated between affinity chromatography, ion-exchange HPLC, and immunological methods.

Statistical methods applied

The result data of the participants for the respective EQA scheme measurand were recalculated by z-scoring. The z-values are the numerical values of the positive or negative standard deviations from the respective CV (for #800) or RMV (for #100 and #145). The resulting z-value ranges of the EQA participants give an impression of the scatter of the individual measured values and are depicted as box-and-whisker plots. The middle line represents the median, whereas the x in the box represents the mean of the z-values. The box includes the lower and the upper quartiles (25%–75%). The whiskers show the minimum and maximum values (±1.5 × the interquartile range (IQR)). The extremes (below or above ±1.5 × IQR) were excluded, as these reported values were mainly compromised by gross errors (sample mix-up, wrong unit, etc.). However, in Figure 1, the outliers are also shown to illustrate the wide scatter of the individual z-values.

Possibly significant changes in the value range over time and in the number of participants of the respective EQA schemes were then investigated by linear regression analysis (least squares method). The degree of association is represented by the coefficient of determination R^2, measured on a scale ranging from −1 through 0 to +1. Complete correlation between two variables being expressed by either +1 or −1. The significance level for the trend line slope >0 was set to $p < 0.05$. All statistical data were calculated using the Microsoft Excel add-in Abacus 3.0, LABanalytics GmbH, Jena, Germany.

Determination of the analytical performance of glucometers according to the ISO standard 15197:2013

A comparative survey and meta-analysis of publications from 2012 to 2022 was performed. A Medline search in this time frame selected these publications. Search terms were *glucometer OR blood glucose monitoring system OR BGMS OR plasma glucose analysis AND ISO 15197*. In brief, ISO 15197:2013 claims the following minimum requirements: First, at least 95% of blood glucose monitoring system (BGMS) results from three different strip lots

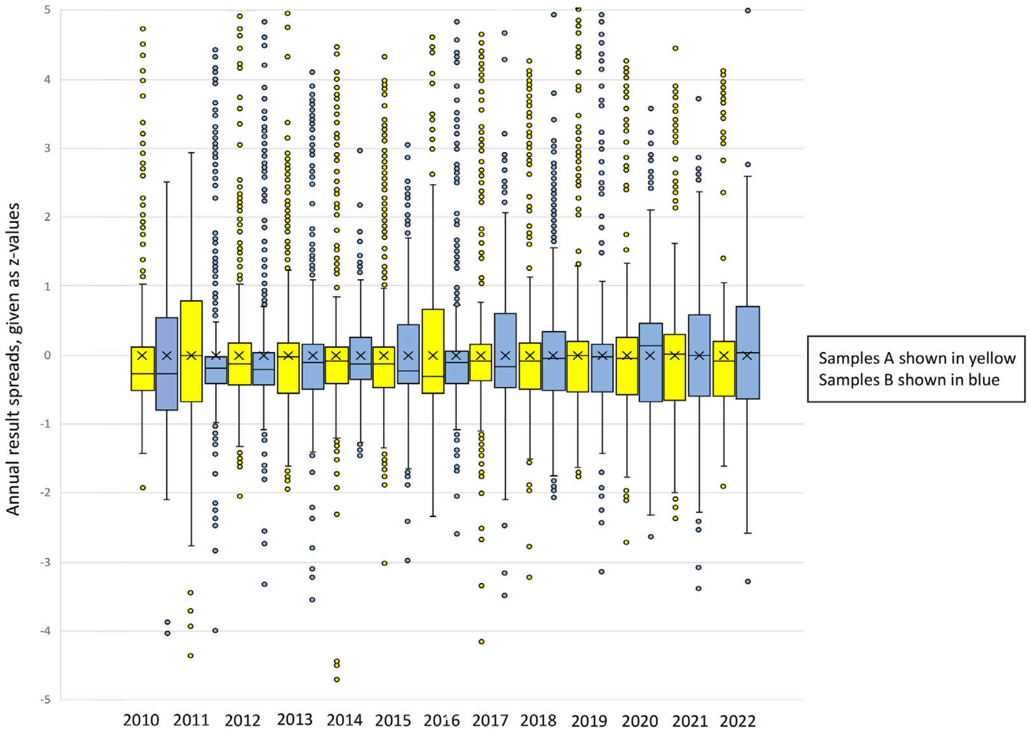


FIGURE 1
Box-and-whisker plots of the annual result spreads for EQA scheme #800 (POCT glucose), given as z-values. Samples A and B are shown for each year. The total sum of results evaluated for this EQA scheme was 17,125. Description of the box-and-whisker plot: The middle line represents the median, whereas the x in the box represents the mean of the z-values. The box includes the lower and the upper quartiles (25%–75%). The whiskers show the minimum and maximum values (±1.5 x interquartile range (IQR)). Single points represent outliers.

have to be within ±15 mg/dL at glucose concentrations <100 mg/dL or within ±15% at ≥100 mg/dL, being compared to a traceable laboratory method. Second, in a consensus (Clarke) error grid analysis, at least 99% of results must be within zones A and B. The different authors checked these requirements.

Results

Prevalence and redistribution of undiagnosed people with diabetes in Germany

General prevalence data

Using the search terms, 122 hits in Medline were found. Most informative for understanding the epidemiological situation in Germany were the annual health reports of the DDG since 2010. For years, these health reports (analysis period 2010–2022) have noted an increasing prevalence of diabetes in the German population. As a result, the number of people with type 2 diabetes in Germany in 2022 rose to approximately 8.7 million (Tönnies et al., 2019); the number of unreported cases could be estimated at two million (Deutsche Diabetes Gesellschaft, 2023). By comparison, the number of diagnosed diabetes cases in 2015 was seven million.

This increase in diabetes prevalence in Germany was predominantly accounted for by subjects aged 65 years and older and those with low educational status, a high body mass index (>30 kg/m^2), and a low physical activity profile (Heidemann et al., 2016). The authors additionally pointed out that the life expectancy of persons with diabetes might have increased more in the last 2 decades than the general population due to more responsive health systems in diagnostics and care (Tönnies et al., 2021). Other potential causes for the increase in prevalence include earlier identification of affected patients using the laboratory parameter HbA1c in whole blood for diagnosis and the increased clinical use of diabetes risk scores (Heidemann et al., 2016). Together with the observation that there has been a decrease in undetected diabetes since 2012, these results suggest that there has been a redistribution from undetected diabetes to diagnosed diabetes in recent years (Deutsche Diabetes Gesellschaft, 2020).

Screening of risk factors and risk scores for type 2 diabetes

To identify patients at increased risk for type 2 diabetes, screening of asymptomatic individuals based on risk factors or risk scores has been recommended in the practice recommendations of the DDG for years (Heidemann and Scheidt-Nave, 2017): Detecting prediabetes based on fasting plasma glucose, 2-h oral glucose tolerance test (OGTT) plasma glucose, or HbA1c. Although the benefits of earlier diagnosis of diabetes are still somewhat unclear, the benefits of lifestyle intervention in individuals with prediabetes have led the US

Preventive Services Task Force to recommend screening for prediabetes and type 2 diabetes in nonpregnant adults 35–70 years of age who are overweight or obese (Goffrier et al., 2017; US Preventive Services Task Force, 2021). The task force concludes with moderate certainty that screening for prediabetes and type 2 diabetes and offering or referring patients with prediabetes to effective preventive interventions has a moderate net benefit.

An indirect approach to determining the risk of diabetes is using risk scores. These allow the estimation of the statistical probability that a person will develop type 2 diabetes in a defined period. Prognostically relevant risk scores allow quantification of risk using a combination of multiple risk parameters and can assist in accurately determining the disease risk for individuals (Deutsche Diabetes Gesellschaft, 2023; Lind et al., 2013).

Gestational diabetes

Since 2012, the German maternity guidelines have recommended a systematic screening program for gestational diabetes using an OGTT (Schäfer-Graf et al., 2018). One of the prerequisites for an effective laboratory screening program is an accurate and precise determination of the glucose concentration in venous plasma. Whenever it is impossible to rapidly test the glucose concentration from whole blood, stabilized blood has to be sent to the analyzing laboratory. The progress made within the last decade was the finding that an effective pre-analytical glycolysis inhibition can only be achieved by using sodium fluoride combined with an acidic citrate buffer (Gambino et al., 2009). Without citrate buffering in the blood collection tubes, false low glucose concentrations may occur, which leads to undetected gestational diabetes.

Results from EQA schemes concerning the total number of participants and the spread of result variations for glucose and HbA1c

Number of participants

In Germany, INSTAND and RfB offer EQA schemes for glucose in plasma as part of the clinical chemistry panel or separate as POCT glucose samples. For the second diabetes measurand, HbA1c, both organizations use fresh whole blood samples with target values assigned with the IFCC reference measurement procedure since 2015.

In 2010–2022, the POCT glucose EQA schemes significantly increased participation in both EQA organizations. In contrast, in the HbA1c dedicated EQA schemes, a substantial increase in the number of participants could be observed only for RfB. For INSTAND, the number of participants decreased between 2015 and 2016. This was due to the shift of the sample matrix from processed to fresh whole blood and the resulting changes in the delivery of samples. Results are summarized in Table 2 and can be retrieved from the Supplementary Figures S1A–D.

Spread of variations of the EQA results

For a better understanding, Figure 1 portrays the spread of result variations throughout the years for the EQA scheme #800 glucose POCT. The outliers are also shown here to illustrate the wide range of the individual z-values. The total sum of results evaluated for this EQA scheme was 17,125.

We found that no significant narrowing of the value spreads could be observed for the glucose EQA schemes #100 and #800. The correlations found showed no significant positive or negative slope. Interestingly, the spreading width of the glucose results in #100 was constantly lower than the width of results in #800. This testifies to a constant high quality of the laboratory analysis in Germany.

The situation is different for the measurand HbA1c. Here, the samples given out by INSTAND have been commutable since 2015, when a new whole blood sample matrix was introduced. In the shorter observation period 2015–2022, significant decreases of result scattering could be observed for two of the three different methods for the analysis of HbA1c: ion-exchange HPLC and immunological methods. Only the affinity chromatography method showed no significant negative slope. The respective statistical data can be found in Table 3.

A complete set of the result spreads for the EQA schemes, shown as box-and-whisker plots, can be retrieved from Supplementary Figures S2–S6.

Data from international studies concerning the analytical performance of glucometers

Sufficiently robust BGMS are a prerequisite for appropriate and safe blood glucose self-monitoring in patients with diabetes. The measurement accuracy of glucometer devices significantly impacts the quality of clinical care and therapy adjustment for these patients (Jendrike et al., 2019). It can be regarded as proven that more significant errors in SMBG devices lead to greater predicted risks of undetected hypoglycemia (Breton and Kovatchev, 2010). Strict accuracy criteria are therefore mandatory for a premarket evaluation to ensure the measurement quality of BGMS systems. These criteria are defined in the international standard ISO 15197 ("*In vitro* diagnostic test systems—Requirements for blood-glucose monitoring systems for self-testing in managing diabetes mellitus"). This standard, first published in 2003 (International Organization for Standardization, 2003), calls for several quality requirements, among them analytical performance evaluations, to guarantee safe and reliable glucose measurements.

Regarding analytical performance, requirements on the accuracy of the respective system (device plus glucose strips) are described in detail, including evaluation design and minimum accuracy criteria. In 2013, a revised version of the norm was published with significant changes like additional stringent accuracy criteria and changes in the testing procedure (International Organization for Standardization, 2013). These criteria were already described in Materials and Methods.

Between 2010 and 2020, a series of methodological studies deals with the compliance of various glucometer systems with the ISO 15197 criteria. Using the search terms, we found 191 hits in Medline; 12 studies were adequate to answer our question. Table 4 displays the percentages of tested reagent system lots that fulfilled the current ISO norm 15197:2013 system accuracy criteria. It can be stated that the percentages increased continuously. This can also be seen in Figure 2. The regression line has an R^2 of 0.2406. Additionally, in an extensive literature review, 58 studies with 143 different SMBG systems between the years 2010 and 2017 were evaluated for

TABLE 2 Linear regression analysis concerning the number of participants in the EQA schemes conducted by INSTAND and RfB for POCT glucose and HbA1c.

Measurand	EQA organization	Number of investigated years (2010–2022)	R^2	p
POCT glucose	RfB	13	0.932	<0.01
POCT glucose	INSTAND	13	0.664	<0.01
HbA1c	RfB	13	0.614	<0.01
HbA1c	INSTAND	13	0.527[a]	–

[a]R = −0.726. Not evaluable due to a change of the offered EQA, material.

TABLE 3 Linear regression analysis concerning the calculated z-values, summarized for samples A and B, representing the variability of the individual results in the INSTAND EQA schemes for glucose (#100 and #800) and HbA1c (#145).

Measurand	Number of investigated years (2010/2015–2022)	R^2	p
Glucose clinical chemistry (#100)	13	0.007	0.689
Glucose POCT (#800)	13	0.066	0.205
HbA1c (#145) affinity chromatography	8	0.120	0.188
HbA1c (#145) ion-exchange HPLC	8	0.245	0.051
HbA1c (#145) immunoassay methods	7[a]	0.369	0.021

[a]The year 2015 was excluded as the number of participants was exceptionally low due to the change in the EQA, material supplied.

TABLE 4 Publications showing percentages of tested reagent system lots that fulfill system accuracy criteria of ISO 15197:2013.

Study author, publication year	Number of tested devices/lots	% Fulfillment of ISO 15197:2013
Baumstark et al. (2012)	20	45
Freckmann et al. (2012)	34	53
Brazg et al. (2013)	21	29
Hasslacher et al. (2014)	27	48
Link et al. (2015)	12	84
Freckmann et al. (2015)	27	78
Yu-Fei et al. (2017)	19	21
Baumstark et al. (2017)	18	83
Jendrike et al. (2018)	12	75
Klonoff et al. (2018)	18	33
Pleus et al. (2020)	18	78
Pleus et al. (2022)	4	100

accuracy. It was shown that newer meters were more likely to pass the ISO 15197:2013 standards (King et al., 2018).

Discussion

German EQA results concerning analytical quality for glucose and HbA1c and analytical performance of glucometers

The survey for both glucose EQA schemes showed no significant change in the spread of result variations over 13 years. For the HbA1c survey, however, there was a significant tendency towards narrowed result spreads, which could be seen in the two methods with the highest number of participants (ion-exchange HPLC and immunological methods), where the affinity chromatography method showed no significant change over time.

During the observation, the POCT glucose EQA schemes showed significant increases in participating laboratories and diabetes-specialized ambulances in both EQA organizations. In contrast, in the HbA1c dedicated EQA schemes, a significant increase in the number of participants could be observed only for RfB. This can be seen as a sign of a consistently good analytical quality of laboratory diagnostics throughout Germany, which

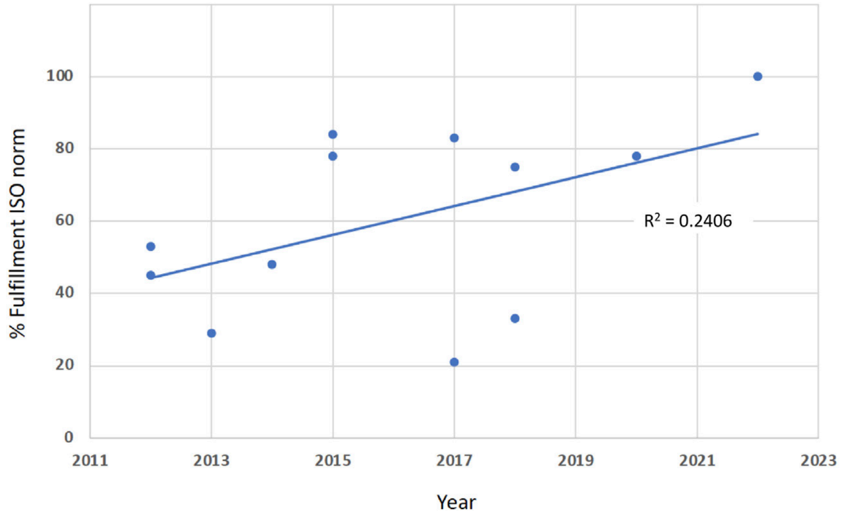


FIGURE 2
The percentages of fulfillment of the EN ISO norm 15197 (*In vitro* diagnostic test systems - Requirements for blood-glucose monitoring systems for self-testing in managing diabetes mellitus), found in various published studies, are increasing between 2012 and 2022 (R^2 = 0.2406).

helps clinicians improve diagnostic and follow-up strategies for patients with diabetes.

The findings for the glucose EQA surveys #100 and #800, however, must be seen against the background that the samples delivered by the EQA organization still suffer from a specimen stability challenge (Wang et al., 2020). The reason is the instability of fresh blood samples, leading to the favored delivery of stabilized sample matrices. In particular, for #800, the matrix effects of such stabilized samples could result in substantial differences in results between the different POCT systems. Therefore, the EQA evaluation can only be carried out according to the consensus value (CV) and not according to the reference method value (RMV) mode.

As depicted in Table 4, the percentages of tested reagent systems that fulfilled the system accuracy criteria of the EN ISO norm 15,197: 2013 increased continuously within the last 2 decades. This testifies to a better analytical quality of glucose measurements within the framework of BGMS and may also be linked to a better diagnosis of previously unknown affected patients. However, it must be mentioned here that the published devices do not necessarily reflect the entire IVD market.

Nevertheless, hypothetical patient scenarios (Eichenlaub et al., 2023) can convey to healthcare professionals and patients a novel understanding of the clinical impact of BGMS accuracy. Despite the standardization of accuracy assessment procedures and requirements, the reliability of the BGMS can still be improved to prevent any adverse clinical events. These are, for example, delayed therapy adjustment, hyperglycemia due to excessive food intake, ketoacidosis, and hypoglycemia due to overcorrection.

Another important point that should be mentioned when evaluating the analytical performance of BGMS is that the reference measurement procedures used for comparison in studies have a considerable impact on the resulting measurement accuracy of BGMS. Since there are systematic differences between the manufacturers' reference measurement procedures used for BGMS calibration and accuracy assessment, this may have potential implications for therapy for patients with diabetes. Therefore, further harmonization of test procedures is desired by various authors to continue the encouraging trend of ever-improving diagnostic capabilities (Freckmann et al., 2022).

What we expected for our EQA HbA1c results, corresponds to the international EurA1c study from 2018: Concerning the analytical performance of HbA1c measurements, this study examined the analytical quality for HbA1c in 2,166 laboratories in 17 different European countries (EurA1c Trial Group, 2018). The results were evaluated according to the criteria of the IFCC model for analytical quality targets. There were two groups in the study. One group received fresh whole-blood samples, and the other lyophilized hemolysate samples. Only one of 20 participating laboratories did not meet the IFCC criterion. Substantial differences between countries and between manufacturer groups were seen by the study group. Germany was in the group with fresh whole-blood samples and achieved a very good result with an IFCC bias of −0.2. Overall, there were no major differences between the fresh whole-blood group and the group using lyophilized hemolysate samples. The findings are in accordance with our results, showing consistently good accuracy of the different HbA1c methods over the entire observation period.

Systematic screening for gestational diabetes—situation since 2012

The German maternity guidelines recommend systematic screening for gestational diabetes using an oral glucose tolerance test since 2012 (Schäfer-Graf et al., 2018). As a result, the prevalence of gestational diabetes significantly increased from 4.6% to 6.8% (2018: 51,318 cases) from 2013 to 2018 (Reitzle et al., 2021). The number continues to grow until 2021 with a prevalence of 8.5%, equivalent to several 63,000 cases. This increase can be explained by several factors: First, in the rise in the age of pregnant women; secondly, by an increase of the pre-conceptional body mass index of the fertile group of women; and finally, by the screening effort itself as a health insurance benefit (Deutsche Diabetes Gesellschaft, 2023). However, laboratory diagnostics also made its contribution. Therefore, analytical aspects are worth mentioning here: To avoid false negative glucose results due to a pronounced metabolic breakdown of the measurand in whole blood (Gambino et al., 2009), a national guideline recommends analyzing the glucose concentration immediately from freshly drawn venous blood by use of quality-assured POCT devices or to use citrate-buffered NaF-tubes when the samples have to be shipped to a laboratory site (Neumaier et al., 2015). This has most likely a positive impact on the false negative results.

Possible link between reduced diabetes mortality, better glycemic control, and an increase in diabetes prevalence by improved laboratory analytics?

Diabetes is a common cause of increased mortality. A recent study on more than 50,000 Spanish individuals impressively showed again that diabetes is associated with a higher premature mortality rate from cardiovascular disease, cancer, and noncardiovascular non-cancer causes compared with the general population (Baena-Díez et al., 2016). Against this background, Chen et al. (Chen et al., 2020) investigated the link between the mortality rate and the prevalence of diabetes mellitus in Caucasian populations. The authors found a significant decline in all-cause mortality since 2000. They concluded that this falling mortality would likely lead to an increasing prevalence despite a stable or even declining incidence of diabetes. They discuss the same public health-related factors as mentioned above, which reduce mortality risk factors. Among them is the optimization of the quality of the analytical techniques with improvements in glycaemic control: Better analytics leads additionally to a higher redistribution rate of undiagnosed diabetes.

Heidemann et al. (Heidemann et al., 2016) further explained factors for the rising prevalence of diabetes in Germany. The authors listed several factors that may be jointly responsible for this observed shift: Increased life expectancy in people with diabetes compared to the general population and the broad clinical application of risk score protocols. Additionally, the drawdown of the cut-off value for the fasting glucose concentration, as proposed by the American Diabetes Association (ADA) in 1997 (Expert Committee on the Diagnosis, 1997) and followed by the WHO 2 years later, combined with the introduction of HbA1c as a valid diagnostic parameter potentially contributed to the observed earlier diabetes diagnosis. However, it must be stated that the diagnostic application of this measurand requires optimized laboratory analytics in terms of accuracy.

Another possible link between better glycemic control of patients with diabetes and optimized laboratory analysis of HbA1c by use of EQA schemes can be deduced from a study by Tollånes (Tollånes et al., 2020). The combination of validated patient data and EQA data showed that patients in offices of general practitioners who participate in HbA1c EQA surveys have lower HbA1c levels. The authors conclude that accurate HbA1c results may improve the diabetes care of the affected patients.

A further Norwegian study investigated various factors that can lead to over- and undertreatment of hyperglycemia. The study examined 10,233 individuals with type 2 diabetes. It was found that a total of 4.1% of patients were potentially overtreated, whereas 7.8% were potentially undertreated, and 11% did not receive an HbA1c measurement (Tran et al., 2021).

Since POCT methods are already widespread in Europe, proficiency testing helps to enhance the quality of the used devices. In particular, for POCT methods measuring HbA1c, there is still room for improvement. A study by Lenters-Westra et al. (2014) showed that not all HbA1c POCT devices met the generally accepted performance criteria. In order to assess the quality class of new POCT devices, efforts should be undertaken for an IFCC standardized comparison method, the adaptation of performance to clinical conditions, and an obligation to register and participate in EQA for proof of quality and quality assurance (Lenters-Westra and English, 2019).

Another study on glucose POCT showed that participants being rated as "failed" in an EQA distribution changed devices more frequently and were, therefore, able to subsequently achieve better analytical results (Bietenbeck et al., 2018).

Conclusion and outlook

Even if the retrospective data analysis only indicates, there appears to be a correlation between lower diabetes mortality, better glycemic control, and increased diabetes prevalence in Germany and consistently high-quality laboratory analytics. This might help to attenuate the high burden of diabetes in terms of its adverse health effects on those affected, but also in terms of its economic impacts on the global healthcare systems. Our assessment of EQA data over time can also be a valuable tool for monitoring the analytical quality of clinical chemistry parameters. It might help to raise the awareness of laboratory professionals for quality concerns.

Good laboratory diagnostics reduce the morbidity and mortality of diseased patients. Yet, diabetes monitoring technology is still on the rise. It is becoming increasingly indicated that patients with insulin-depending diabetes use CGM systems nowadays. The American Association of Clinical Endocrinology Clinical Practice Guideline from 2021 recommends the use of advanced technology in the management of people with diabetes to effectively achieve the glycemic targets, thereby improving quality and convenience of life, reducing the burden of care, and offering a personalized approach to self-testing (Grunberger et al., 2021). However, quality assurance measures comparable to the EQA protocols described in this study still need to be established internationally.

Laboratory diagnostics can also help detect patients with slowly progressive late-onset autoimmune diabetes in adults (LADA). Anti-islet autoantibodies to insulin (IAA), glutamic acid decarboxylase (GADA), tyrosine phosphatase-like protein IA-2 (IA-2A), and zinc transporter 8 (ZnT8A) are currently employed in the improved diagnostic process (Kawasaki, 2023). Here, too, EQA programs have already been established in Germany by both accredited EQA organizations.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The data of the investigated EQA schemes are provided by INSTAND as excel sheets and are stored in the institute. Requests to access these datasets should be directed to p.luppa@tum.de.

Author contributions

PL: Conceptualization, Data curation, Formal Analysis, Methodology, Supervision, Writing–original draft, Writing–review and editing. MZ: Data curation, Formal Analysis, Writing–original draft. MP: Investigation, Software, Writing–review and editing. PK: Conceptualization, Data curation, Formal Analysis, Writing–review and editing. NW: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Writing–original draft. LV: Writing–review and editing. GF: Conceptualization, Data curation, Investigation, Methodology, Supervision, Validation, Writing–original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The presented study had no dedicated funding and was performed as part of the authors' professional employment.

Acknowledgments

The authors thank Prof. Kurt Ulm and PD Markus Thaler from Munich for statistical advice. We are also grateful to Delia Waldenmaier, Ulm, for the critical proofreading of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2024.1371426/full#supplementary-material

References

Baena-Díez, J. M., Peñafiel, J., Subirana, I., Ramos, R., Elosua, R., Marín-Ibañez, A., et al. (2016). Risk of cause-specific death in individuals with diabetes: a competing risks analysis. *Diabetes Care* 39 (11), 1987–1995. doi:10.2337/dc16-0614

Baumstark, A., Jendrike, N., Pleus, S., Haug, C., and Freckmann, G. (2017). Evaluation of accuracy of six blood glucose monitoring systems and modeling of possibly related insulin dosing errors. *Diabetes Technol. Ther.* 19, 580–588. doi:10.1089/dia.2016.0408

Baumstark, A., Pleus, S., Schmid, C., Link, M., Haug, C., and Freckmann, G. (2012). Lot-to-lot variability of test strips and accuracy assessment of systems for self-monitoring of blood glucose according to ISO 15197. *J. Diabetes Sci. Technol.* 6, 1076–1086. doi:10.1177/193229681200600511

Bietenbeck, A., Geilenkeuser, W. J., Klawonn, F., Spannagl, M., Nauck, M., Petersmann, A., et al. (2018). External quality assessment schemes for glucose measurements in Germany: factors for successful participation, analytical performance and medical impact. *Clin. Chem. Lab. Med.* 56, 1238–1250. doi:10.1515/cclm-2017-1142

Blasutig, I. M., Wheeler, S. E., Bais, R., Dabla, P. K., Lin, J., Perret-Liaudet, A., et al. (2023). External quality assessment practices in medical laboratories: an IFCC global survey of member societies. *Clin. Chem. Lab. Med.* 61, 1404–1410. doi:10.1515/cclm-2023-0057

Brazg, R. L., Klaff, L. J., and Parkin, C. G. (2013). Performance variability of seven commonly used self-monitoring of blood glucose systems: clinical considerations for patients and providers. *J. Diabetes Sci. Technol.* 7, 144–152. doi:10.1177/193229681300700117

Breton, M. D., and Kovatchev, B. P. (2010). Impact of blood glucose self-monitoring errors on glucose variability, risk for hypoglycemia, and average glucose control in type 1 diabetes: an *in silico* study. *J. Diabetes Sci. Technol.* 4, 562–570. doi:10.1177/193229681000400309

Brutsaert, E. F. (2023). *Diabetes mellitus (DM) - hormonal and metabolic disorders.* MSD Manual Consumer Version.

Bundesärztekammer, (2023). Aktualisierung der Richtlinie der Bundesärztekammer zur Qualitätssicherung laboratoriumsmedizinischer Untersuchungen - rili-BÄK. *Dtsch. Ärzteblatt* 120, A994. doi:10.3238/arztebl.2023.rili_baek_QS_Labor

Chen, L., Islam, R. M., Wang, J., Hird, T. R., Pavkov, M. E., Gregg, E. W., et al. (2020). A systematic review of trends in all-cause mortality among people with diabetes. *Diabetologia* 63, 1718–1735. doi:10.1007/s00125-020-05199-0

Deutsche Diabetes Gesellschaft (DDG) und DiabetesDE (2020). *Deutsche DIabetesHilfe. Deutscher Gesundheitsbericht Diabetes 2020.* Mainz, Germany: Verlag Kirchheim + Co GmbH.

Deutsche Diabetes Gesellschaft (DDG) und diabetesDE - Deutsche DIabetesHilfe (2023). *Deutscher Gesundheitsbericht Diabetes 2023. Die Bestandsaufnahme*. Mainz, Germany: Verlag Kirchheim + Co GmbH.

Eichenlaub, M., Pleus, S., Shaginian, R., Richardson, J., Pardo, S., Stuhr, A., et al. (2023). Impact of blood glucose monitoring system accuracy on clinical decision making for diabetes management. *J. Diabetes Sci. Technol.* 17, 683–689. doi:10.1177/19322968221080916

EurA1c Trial Group (2018). EurA1c: the European HbA1c trial to investigate the performance of HbA1c assays in 2166 laboratories across 17 countries and 24 manufacturers by use of the IFCC model for quality targets. *Clin. Chem.* 64, 1183–1192. doi:10.1373/clinchem.2018.288795

Expert Committee on the Diagnosis (1997). Expert committee on the diagnosis and classification of diabetes mellitus: report of the expert committee on the diagnosis and classification of diabetes mellitus. *Dia- betes Care* 20, 1183–1197. doi:10.2337/diacare.20.7.1183

Freckmann, G., Baumstark, A., Jendrike, N., Mende, J., Schauer, S., Link, M., et al. (2022). Impact of two different reference measurement procedures on apparent system accuracy of 18 CE-marked current-generation blood glucose monitoring systems. *J. Diabetes Sci. Technol.* 16, 1076–1088. doi:10.1177/1932296820948873

Freckmann, G., Link, M., Schmid, C., Pleus, S., Baumstark, A., and Haug, C. (2015). System accuracy evaluation of different blood glucose monitoring systems following

ISO 15197:2013 by using two different comparison methods. *Diabetes Technol. Ther.* 17, 635–648. doi:10.1089/dia.2015.0085

Freckmann, G., Schmid, C., Baumstark, A., Pleus, S., Link, M., and Haug, C. (2012). System accuracy evaluation of 43 blood glucose monitoring systems for self-monitoring of blood glucose according to DIN EN ISO 15197. *J. Diabetes Sci. Technol.* 6, 1060–1075. doi:10.1177/193229681200600510

Gambino, R., Piscitelli, J., Ackattupathil, T. A., Theriault, J. L., Andrin, R. D., Sanfilippo, M. L., et al. (2009). Acidification of blood is superior to sodium fluoride alone as an inhibitor of glycolysis. *Clin. Chem.* 55, 1019–1021. doi:10.1373/clinchem.2008.121707

Goffrier, B., Schulz, M., and Bätzing-Feigenbaum, J.: (2017). Administrative Prävalenzen und Inzidenzen des Diabetes mellitus von 2009 bis 2015. Versorgungsatlas-Bericht 2017. 17/03: doi:10.20364/VA-17.03

Grunberger, G., Sherr, J., Allende, M., Blevins, T., Bode, B., Handelsman, Y., et al. (2021). American association of clinical Endocrinology clinical practice guideline: the use of advanced technology in the management of persons with diabetes mellitus. *Endocr. Pract.* 27, 505–537. doi:10.1016/j.eprac.2021.04.008

Hasslacher, C., Kulozik, F., and Platten, I. (2014). Analytical performance of glucose monitoring systems at different blood glucose ranges and analysis of outliers in a clinical setting. *J. Diabetes Sci. Technol.* 8, 466–472. doi:10.1177/1932296814522804

Heidemann, C., Du, Y., Paprott, R., Haftenberger, M., Rathmann, W., and Scheidt-Nave, C. (2016). Temporal changes in the prevalence of diagnosed diabetes, undiagnosed diabetes and prediabetes: findings from the German health interview and examination surveys in 1997-1999 and 2008-2011. *Diabet. Med.* 33, 1406–1414. doi:10.1111/dme.13008

Heidemann, C., and Scheidt-Nave, C. (2017). Prevalence, incidence and mortality of diabetes mellitus in adults in Germany - a review in the framework of the diabetes surveillance. *J. Health Monit.* 2, 98–121. doi:10.17886/RKI-GBE-2017-062

International Organization for Standardization. *In vitro* diagnostic test systems-requirements for blood-glucose monitoring systems for self-testing in managing diabetes mellitus. EN ISO 15197:2003.

International Organization for Standardization. *In vitro* diagnostic test systems -requirements for blood-glucose monitoring systems for self-testing in managing diabetes mellitus. EN ISO 15197:2013.

Jendrike, N., Baumstark, A., Pleus, S., Liebing, C., Beer, A., Flacke, F., et al. (2018). Evaluation of four blood glucose monitoring systems for self-testing with built-in insulin dose advisor based on ISO 15197:2013: system accuracy and hematocrit influence. *Diabetes Technol. Ther.* 20, 303–313. doi:10.1089/dia.2017.0391

Jendrike, N., Baumstark, A., Pleus, S., Mende, J., Haug, C., and Freckmann, G. (2019). Assessment of system accuracy, intermediate measurement precision, and measurement repeatability of a blood glucose monitoring system based on ISO 15197. *J. Diabetes Sci. Technol.* 13, 235–241. doi:10.1177/1932296818821105

Kawasaki, E. (2023). Anti-islet autoantibodies in type 1 diabetes. *Int. J. Mol. Sci.* 24, 10012. doi:10.3390/ijms241210012

King, F., Ahn, D., Hsiao, V., Porco, T., and Klonoff, D. C. (2018). A review of blood glucose monitor accuracy. *Diabetes Technol. Ther.* 20, 843–856. doi:10.1089/dia.2018.0232

Klonoff, D. C., Parkes, J. L., Kovatchev, B. P., Kerr, D., Bevier, W. C., Brazg, R. L., et al. (2018). Investigation of the accuracy of 18 marketed blood glucose monitors. *Diabetes Care* 41, 1681–1688. doi:10.2337/dc17-1960

Kravarusic, J., and Aleppo, G. (2020). Diabetes technology use in adults with type 1 and type 2 diabetes. *Endocrinol. Metab. Clin. North Am.* 49, 37–55. doi:10.1016/j.ecl.2019.10.006

Lenters-Westra, E., and English, E. (2019). Analysis: investigating the quality of POCT devices for HbA1c, what are our next steps? *J. Diabetes Sci. Technol.* 13, 1154–1157. doi:10.1177/1932296819850838

Lenters-Westra, E., and Slingerland, R. J. (2014). Three of 7 hemoglobin A1c point-of-care instruments do not meet generally accepted analytical performance criteria. *Clin. Chem.* 60, 1062–1072. doi:10.1373/clinchem.2014.224311

Lin, R., Brown, F., James, S., Jones, J., and Ekinci, E. (2021). Continuous glucose monitoring: a review of the evidence in type 1 and 2 diabetes mellitus. *Diabet. Med.* 38, e14528. doi:10.1111/dme.14528

Lind, M., Garcia-Rodriguez, L. A., Booth, G. L., Cea-Soriano, L., Shah, B. R., Ekeroth, G., et al. (2013). Mortality trends in patients with and without diabetes in ontario, Canada, and the UK from 1996 to 2009: a population-based study. *Diabetologia* 56, 2601–2608. doi:10.1007/s00125-013-3063-1

Link, M., Schmid, C., Pleus, S., Baumstark, A., Rittmeyer, D., Haug, C., et al. (2015). System accuracy evaluation of four systems for self-monitoring of blood glucose following ISO 15197 using a glucose oxidase and a hexokinase-based comparison method. *J. Diabetes Sci. Technol.* 9, 1041–1050. doi:10.1177/1932296815580161

Neumaier, M., Luppa, P. B., Koschinsky, T., Siegel, E., Freckmann, G., and Heinemann, L. (2015). Updated requirements for measurement quality and quality assurance of point-of-care testing (POCT) - blood glucose measurement systems with unit-use reagents suitable for the initial diagnosis of diabetes manifested in pregnancy or gestational diabetes mellitus (GDM) according to the GDM guideline of the German diabetes association (DDG). *Diabetologie* 10, 1–3. doi:10.1055/s-0035-1553622

Petersmann, A., Macdonald, R., and Nauck, M. (2022). Disregarded measurement uncertainty contributions and their magnitude in measuring plasma glucose. *J. Diabetes Sci. Technol.* 16, 161–167. doi:10.1177/1932296820966353

Pleus, S., Baumstark, A., Jendrike, N., Mende, J., Link, M., Zschornack, E., et al. (2020). System accuracy evaluation of 18 CE-marked current-generation blood glucose monitoring systems based on EN ISO 15197:2015. *BMJ Open Diabetes Res. Care* 8, e001067. doi:10.1136/bmjdrc-2019-001067

Pleus, S., Baumstark, A., Schauer, S., Kölle, J., Jendrike, N., Mende, J., et al. (2022). User performance evaluation and system accuracy assessment of four blood glucose monitoring systems with color coding of measurement results. *J. Diabetes Sci. Technol.*, 193229682211419. doi:10.1177/19322968221141926

Reitzle, L., Schmidt, C., Heidemann, C., Icks, A., Kaltheuner, M., Ziese, T., et al. (2021). Gestationsdiabetes in Deutschland: zeitliche Entwicklung von Screeningquote und Prävalenz. *J. Health Monit.* 6, 3–19.

Schäfer-Graf, U. M., Gembruch, U., Kainer, F., Groten, T., Hummel, S., Hösli, I., et al. (2018). Gestational diabetes mellitus (GDM) - diagnosis, treatment and follow-up. Guideline of the DDG and DGGG (S3 level, AWMF registry number 057/008, february 2018). *Geburtshilfe Frauenheilkd* 78, 1219–1231. doi:10.1055/a-0659-2596

Schleicher, E., Gerdes, C., Petersmann, A., Müller-Wieland, D., Müller, U. A., Freckmann, G., et al. (2022). Definition, classification and diagnosis of diabetes mellitus. *Exp. Clin. Endocrinol. Diabetes* 130, S1–S8. doi:10.1055/a-1624-2897

Tollånes, M. C., Jenum, A. K., Berg, T. J., Løvaas, K. F., Cooper, J. G., and Sandberg, S. (2020). Availability and analytical quality of hemoglobin A1c point-of-care testing in general practitioners' offices are associated with better glycemic control in type 2 diabetes. *Clin. Chem. Lab. Med.* 58, 1349–1356. doi:10.1515/cclm-2020-0026

Tönnies, T., Baumert, J., Heidemann, C., von der Lippe, E., Brinks, R., and Hoyer, A. (2021). Diabetes-free life expectancy and years of life lost associated with type 2 diabetes: projected trends in Germany between 2015 and 2040. *Popul. Health Metr.* 19, 38. doi:10.1186/s12963-021-00266-z

Tönnies, T., Röckl, S., Hoyer, A., Heidemann, C., Baumert, J., Du, Y., et al. (2019). Projected number of people with diagnosed type 2 diabetes in Germany in 2040. *Diabet. Med.* 36, 1217–1225. doi:10.1111/dme.13902

Tran, A. T., Berg, T. J., Mdala, I., Gjelsvik, B., Cooper, J. G., Sandberg, S., et al. (2021). Factors associated with potential over- and undertreatment of hyperglycaemia and annual measurement of HbA1c in type 2 diabetes in Norwegian general practice. *Diabet. Med.* 38, e14500. doi:10.1111/dme.14500

US Preventive Services Task Force; Davidson, K. W., Barry, M. J., Mangione, C. M., Cabana, M., Caughey, A. B., and Davis, E. M. (2021). Screening for prediabetes and type 2 diabetes: US preventive Services task force recommendation statement. *JAMA* 326, 736–743. doi:10.1001/jama.2021.12531

Wang, Y., Plebani, M., Sciacovelli, L., Zhang, S., Wang, Q., and Zhou, R. (2020). Commutability of external quality assessment materials for point-of-care glucose testing using the clinical and laboratory standards Institute and international federation of clinical chemistry approaches. *J. Clin. Lab. Anal.* 34, e23327. doi:10.1002/jcla.23327

Weykamp, C. (2013). HbA1c: a review of analytical and clinical aspects. *Ann. Lab. Med.* 33, 393–400. doi:10.3343/alm.2013.33.6.393

Yu-Fei, W., Wei-Ping, J., Ming-Hsun, W., Miao, O. C., Ming-Chang, H., Chi-Pin, W., et al. (2017). Accuracy evaluation of 19 blood glucose monitoring systems manufactured in the asia-pacific region: a multicenter study. *J. Diabetes Sci. Technol.* 11, 953–965. doi:10.1177/1932296817705143

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 20 June 2024
DOI 10.3389/fmolb.2024.1401619

Check for updates

OPEN ACCESS

EDITED BY
Matteo Becatti,
University of Firenze, Italy

REVIEWED BY
Irina Nakashidze,
Shota Rustaveli State University, Georgia
Joško Osredkar,
University Medical Centre Ljubljana, Slovenia

*CORRESPONDENCE
Stefan Holdenrieder,
✉ s.holdenrieder@tum.de

RECEIVED 15 March 2024
ACCEPTED 30 May 2024
PUBLISHED 20 June 2024

CITATION
Kremser M, Weiss N, Kaufmann-Stoeck A, Vierbaum L, Schmitz A, Schellenberg I and Holdenrieder S (2024), Longitudinal evaluation of external quality assessment results for CA 15-3, CA 19-9, and CA 125.
Front. Mol. Biosci. 11:1401619.
doi: 10.3389/fmolb.2024.1401619

Longitudinal evaluation of external quality assessment results for CA 15-3, CA 19-9, and CA 125

Marcel Kremser[1], Nathalie Weiss[1], Anne Kaufmann-Stoeck[1], Laura Vierbaum[1], Arthur Schmitz[1], Ingo Schellenberg[1,2] and Stefan Holdenrieder[1,3]*

[1]INSTAND e.V., Society for Promoting Quality Assurance in Medical Laboratories, Duesseldorf, Germany, [2]Center of Life Sciences, Institute of Bioanalytical Sciences (IBAS), Anhalt University of Applied Sciences, Bernburg, Germany, [3]Munich Biomarker Research Center, Institute of Laboratory Medicine, Deutsches Herzzentrum München, Technische Universität München, Munich, Germany

Background: Tumor markers are established laboratory tools that help to diagnose, estimate prognosis, and monitor the course of cancer. For meaningful decision-making in patient care, it is essential that methods and analytical platforms demonstrate high sensitivity, specificity, precision, and comparability. Regular participation at external quality assessment (EQA) schemes is mandatory for laboratories. Here, a longitudinal evaluation of EQA data was performed to assess the performance of tumor marker assays over time.

Methods: Longitudinal data of the cancer antigens (CA) 15-3 (n = 5,492), CA 19-9 (n = 6,802), and CA 125 (n = 5,362) from 14 INSTAND EQAs conducted between 2019 and 2023 were evaluated. A median of 197, 244 and 191 laboratories participated at the EQAs for CA 15-3, CA 19-9 and CA 125, respectively. Data evaluation encompasses intra- and inter-manufacturer specific variations over time, assay precision, and adherence to the EQA limits of ±24% for CA 15-3, ±27% for CA 19–9 and ±36% for CA 125.

Results: The study showed median manufacturer-dependent differences of up to 107% for CA 15-3, 99% for CA 125, and even 549% for CA 19-9 between the highest and the lowest methods over the studied period. Regarding the normalized median of all methods, the values of the most deviant methods were 0.42 for CA 15-3, 7.61 for CA 19-9, and 1.82 for CA 125. Intra-manufacturer variability was generally low, with median coefficients of variation (CV) below 10%. As the methods were evaluated according to method-specific consensus values, most participants passed the EQAs within the acceptance criteria. When the criteria were consistently set at 24%, the central 90% of participants passed the EQAs in 78.6%–100% for CA 15-3 (with exception of AX), 89.3%–100% for CA 125, and 64.3%–100% for CA 19-9.

Conclusion: While intra-method precision of most analytical platforms is acceptable for all three tumor markers, considerable inter-method variability was observed over the whole studied period demonstrating the necessity for better standardization and harmonization of the methods, development of international reference materials, and comprehensive commutability studies with patient samples.

KEYWORDS

external quality assessment, tumor marker, cancer antigen, CA 15-3, CA 19-9, CA 125, EQA, INSTAND

1 Introduction

Cancer remains a challenge to public health worldwide (Bray et al., 2021; Sung et al., 2021). As our understanding of cancer biology continues to advance, so does the need for improved diagnostic tools for the detection, risk assessment, and monitoring of therapeutic responses. Tumor markers have risen in prominence as potential indicators for the presence and progression of cancer (Filella et al., 2023).

Among the diverse array of tumor markers, the cancer antigens (CA) 15-3, CA 19-9, and CA 125 have emerged as useful tools in the detection and management of various cancer entities (Stieber and Heinemann, 2008a). CA 15-3, also known as Mucin-1 (MUC-1), is a 300 kDa carbohydrate antigen found in normal breast and breast cancer cells (Gang et al., 1985; Duffy et al., 2000). It is also expressed by other types of cancer, such as lung cancer and gastric cancer, and appears in elevated levels in the blood serum and plasma of patients with non-cancer-related conditions like cirrhosis, hepatitis and benign breast diseases (Duffy et al., 2010).

CA 19-9, also known as Sialyl Lewis-antigen, is a 36 kDa glycolipid that emerges from the generation of a monoclonal antibody against a colon carcinoma cell line (DelVillano and Zurawski, 1983). Elevated levels of CA 19-9 are notably exhibited in the blood of patients with various malignancies, including gastric, lung, colon and pancreatic cancers (Lee et al., 2020). Furthermore, high levels of CA 19-9 in the blood are observed in non-malignant conditions such as benign pancreatobiliary, hepatic and pulmonary diseases, and in cases of thyroiditis, diabetes mellitus, and autoimmune disorders (Trape et al., 2011; Kim et al., 2020).

CA 125, also known as MUC-16, is a 200 kDa membrane glycoprotein expressed on the surface of ovarian cancer cells (Charkhchi et al., 2020). It is defined by the monoclonal antibody OC125, which is derived from human ovarian cancer cell lines. In addition to ovarian cancer, elevated levels in the blood are found in conjunction with lung, endometrial, pancreatic, breast, and colon cancer, as well as with physiological conditions such as menstruation and pregnancy (Ghosh et al., 2013). Given its susceptibility of being elevated under a range of circumstances, CA 125 is used in combination with other tumor markers, like human epididymis protein 4 (HE4), to assess the risk of suspicious pelvic masses (Moore et al., 2009; Escudero et al., 2011).

Although extensive research has been conducted on these tumor markers, challenges persist in achieving standardization and harmonization across methods (Mongia et al., 2006; Slev et al., 2006; La'ulu and Roberts, 2007; Passerini et al., 2007; Serdarevic, 2018). In 2005, the Society for Promoting Quality Assurance in Medical Laboratories (INSTAND) observed a manufacturer-dependent bias of up to 44% for CA 15-3, 194% for CA 19-9 and 162% for CA 125 as part of external quality assessment (EQA) results (Reinauer and Wood, 2005). INSTAND is accredited according to ISO17043 and is a reference institute of the German Medical Association. It has been conducting EQAs since 1966. Considerable variation has also been reported in clinical studies that compare different manufacturers (Stieber et al., 2008b; Holdenrieder et al., 2008; Molina et al., 2008). This is a matter of concern, as the ability to compare results across laboratories, manufacturers, and platforms is crucial for the meaningful interpretation of clinical data. This is particularly true given that the reference limits of different methods are often similar (La'ulu and Roberts, 2007). If cancer patients undergoing therapy or post-treatment surveillance receive tumor marker results from different laboratories utilizing different methods, the lack of standardization and harmonization can lead to erroneous interpretations of the marker dynamics. Furthermore, EQA providers are required to establish acceptance criteria for method-specific EQA schemes, which are essential for the interpretation of clinically meaningful results. Additionally, they must monitor the performance of analytes and methods over time.

In a recent analysis of EQA data on the current quality of the tumor markers alpha-feto protein (AFP) and carcinoembryonic antigen (CEA), for which there are international reference standard materials, we found a better level of standardization between 2018 and 2022 compared to that reported in 2005 (Wojtalewicz et al., 2023). In this study we performed a longitudinal assessment of EQA data for the tumor markers CA 15-3, CA 19-9, and CA 125 for which international reference standard materials have not yet been developed. We compared intra- and inter-method variations between EQA participants using the most common analytical platforms and tested their adherence to EQA limits.

2 Materials and methods

2.1 Sample materials

The matrix for the EQA samples was composed of human serum pools stabilized with 0.02% sodium azide. No other synthetic substances were added. To reach defined tumor marker concentrations, the matrix was spiked with non-synthetic tumor antigens from respective tumor tissue cell lines. Sample concentrations were selected based on clinical relevance and in accordance with the guidelines of the German Medical Association (RiliBÄK). The manufacturer declared and confirmed the homogeneity and stability of each sample batch. During the EQA surveys, the liquid samples were stored at 2°C–8°C until shipment.

2.2 EQA procedure

The INSTAND EQA scheme for tumor marker detection is conducted six times a year on a global scale. There are no exclusions for participants. For each survey, participating laboratories receive two EQA samples with different concentrations. The laboratories are required to report their quantitative results for CA 15-3, CA 19-9 and CA 125, along with other tumor markers, and provide information to INSTAND about the respective analytical platforms, methods, reagents, and manufacturers. Participating laboratories report this information via the RV-Online platform (https://rv-online.instandev.de).

As there is no available reference method for tumor marker quantification, the consensus value of manufacturer-specific collectives, calculated using algorithm A (ISO13528, (2022), Section C3), serves as the target value for evaluating participant results and laboratory certification. The EQA passing criterion for CA 15-3 is set at ±24% of the consensus value over the entire evaluation period. This is in accordance with the RiliBÄK (Bundesärztekammer, 2023). For CA 19-9 and CA 125, which are not covered in the current guideline, the EQA criteria are set at ±27% and ±36%, respectively.

2.3 Data analysis and statistics

In the present study EQA surveys conducted between January 2019 and May 2023 for the tumor markers CA 15-3, CA 19-9 and CA 125 were examined. As in the previously published tumor marker study (Wojtalewicz et al., 2023), only data from the three annual EQAs with the highest number of participants, namely January, May, and October, were included in the evaluation (Supplementary Table S1). The lower participant number EQA schemes have been excluded due to low statistically significance. In total 14 CA 15-3, CA 19-9 and CA 125 EQA surveys with two samples each were analyzed.

The EQA samples had different concentrations of the tumor markers that mirrored the relevant value range for clinical decision making. This was close to the cut-off values of the so-called reference range (95th percentile of healthy individuals), which is around 30 kU/L for CA 15-3, 35 kU/L for CA 125, and 37 kU/L for CA 19-9 for most manufacturers and methods, and at slightly or strongly elevated levels as often seen in different cancer stages. For better orientation, cut-off values are highlighted with a red line in the figures.

The EQA data were analyzed in a manufacturer-dependent manner (Supplementary Table S2). We focused on manufacturer collectives with a minimum of six participants per survey, resulting in six collectives for the analysis of the CA 15-3 results, seven collectives for CA 19-9, and six collectives for CA 125. These were, in alphabetical order, Abbott (AB), Beckman (BE), bioMérieux (AX), Diasorin (DO), Roche (RO), Siemens (SI), and Tosoh (TH, for CA 19-9 only).

The results were illustrated using combined dot plots and box plot diagrams to visualize the distributions of the values in terms of median, interquartile range, and whiskers and to make them comparable over time.

The SI collective comprised four manufacturer sub-collectives consolidated under Siemens. In some EQA surveys, we observed a multimodality in the SI collective, but to gain a comprehensive understanding of the value distribution, all results from the SI cohort were included in the general box plot analysis. Additionally, the SI collective was divided into subgroups (Supplementary Table S1; Supplementary Figures S1–S4). Due to the multimodality of the SI collective, we specifically presented the normalized median for the more substantial sub-collectives Bayer Health (BG), DPC Biermann (DG) and Siemens Healthineers (SIE).

The collective median of each survey was normalized in relation to the overall median of the respective survey. The coefficients of variation (CVs) were calculated to assess the scatter within the manufacturer collectives; for the SI collective, the three sub-collectives BG, DG and SIE were considered separately.

In a further step, the inter-laboratory performance quality of CA 15-3, CA 19-9 and CA 125 as well as the manufacturer-dependent value distribution were analyzed in relation to the EQA success criteria. Here the central 80% (10th to 90th percentiles) and the central 90% (5th to 95th percentiles) of the participants of each manufacturer were compared to the acceptance criteria of each tumor marker.

We used jmp 17.2.0 from SAS Institute (Cary, NC, United States) for the basic statistical analyses. The overlay images were generated using version 2.10.34 of the Gnu image manipulation software.

3 Results

The data from the 14 EQA surveys, conducted in January, May and October between 2019 and 2023, were examined for the tumor markers CA 15-3, CA 19-9, and CA 125. The participating laboratories collectively provided 5,492 results for CA 15-3, 6,802 results for CA 19-9 and 5,362 results for CA 125. A median of 197 laboratories participated at the EQAs for CA 15-3 (minimum 172, maximum 219), 244 laboratories for CA 19-9 (minimum 214, maximum 275), and 191 laboratories for CA 125 (minimum 165, maximum 220). The detailed numbers of results per manufacturer are displayed in Supplementary Table S2. Regarding outlier management, sample mix-ups or reporting errors resulted in the exclusion of 35 results for CA 15-3, 20 results for CA 19-9, and 16 results for CA 125.

3.1 CA 15-3 EQA results

Notable disparities in concentrations of CA 15-3 were observed across manufacturers, with median variations reaching as high as 107% between BE and SI and the maximum variations reaching as high as 171% between BE and DO. For other methods, the differences were lower as displayed in detail in Supplementary Table S3. The BE collective consistently reported the lowest values and never overlapped with results from other collectives (Figure 1A). In contrast, the SI collective often reported the highest values. Excluding the BE collective from the analysis substantially reduced the highest manufacturer-specific concentration differences to 25%, as seen between SI and AX in the January 2021 survey.

The trend of BE consistently reporting the lowest values became even more apparent when the normalized median differences between manufacturer collectives (Figure 1B) and the median values along with the minimum and maximum values of the normalized median differences for each manufacturer collective

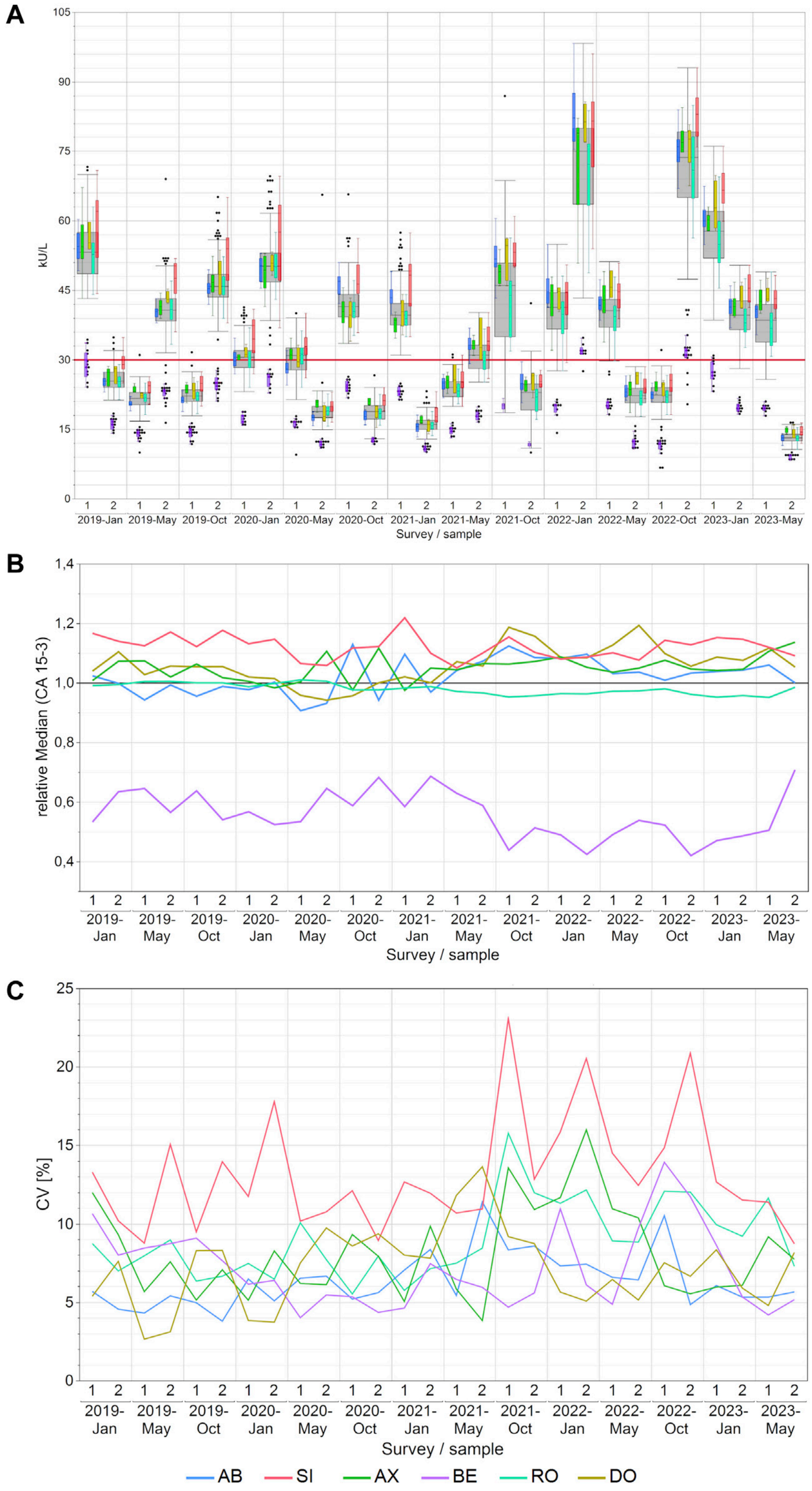


FIGURE 1
Manufacturer-dependent analysis of CA 15-3 EQA results, encompassing an all-results overview **(A)**, comparisons of manufacturer-dependent median differences relative to the overall median **(B)**, and evaluations of manufacturer-dependent CVs **(C)** between 2019 and 2023. Data are presented for two samples per survey. The gray boxes represent all results for the respective sample, while smaller, colored box plots overlay the total results (blue: AB, green: AX, cyan: RO, violet: BE, red: SI, ochre: DO). A red line marks the 30 kU/L cut-off value, and black dots denote outliers excluded from the colored boxes. The whiskers extend from the 1st quartile minus 1.5 times the interquartile range to the 3rd quartile plus 1.5 times the interquartile range.

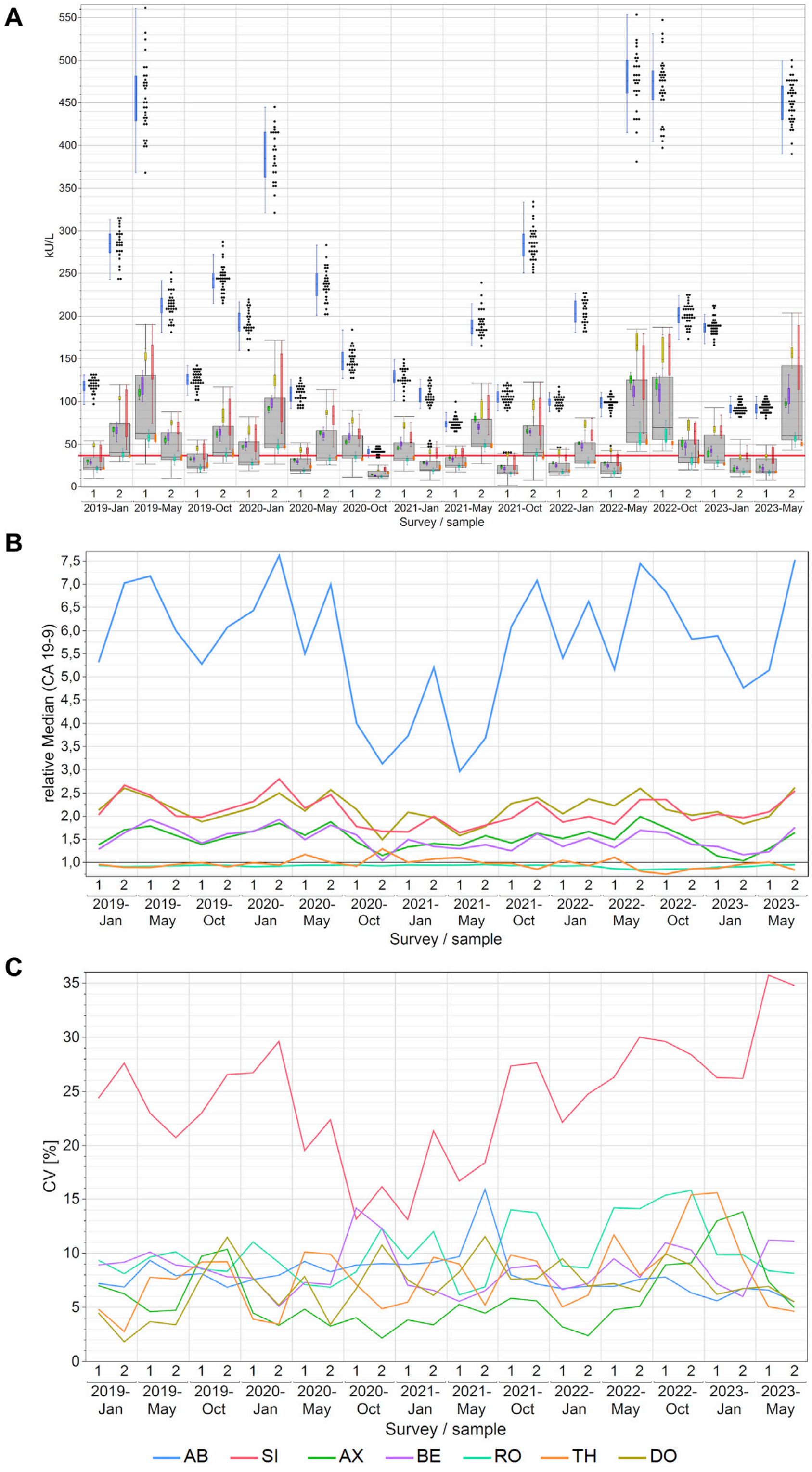


FIGURE 2
Manufacturer-dependent analysis of CA 19-9 EQA results, encompassing an all-results overview **(A)**, comparisons of manufacturer-dependent median differences relative to the overall median **(B)**, and evaluations of manufacturer-dependent CVs **(C)** between 2019 and 2023. Data are presented for two samples per survey. The gray boxes represent all results for the respective sample, while smaller, colored box plots overlay the total results (blue: AB, green: AX, cyan: RO, violet: BE, red: SI, ochre: DO, orange: TH). A red line marks the 37 kU/L cut-off value, and black dots denote outliers excluded from the colored boxes. The whiskers extend from the 1st quartile minus 1.5 times the interquartile range to the 3rd quartile plus 1.5 times the interquartile range.

were considered (Supplementary Table S4). Notably, the BE collective exhibited the lowest relative median value of 0.54—noticeably lower than the other collectives.

The median intra-manufacturer coefficients of variation (CVs) for CA 15-3 measurements mostly remained below 10% (maximum 16%), pointing to a high level of assay precision (Figure 1C; Supplementary Table S5). The only exception to this pattern was the SI collective, which achieved a maximum CV of up to 23%. Subdividing the SI collective into sub-collectives showed lower median CV below 10% (maximum 20% for the SIE subgroup; Supplementary Figure S1A).

3.2 CA 19-9 EQA results

For CA 19-9, the AB collective consistently reported considerably higher values and never overlapped with the other collectives. Its values occasionally reached very high levels of approximately 560 kU/L. This contrasted starkly with other companies, where measurements typically did not exceed 200 kU/L. Conversely, the RO and TH collectives consistently reported the lowest values for CA 19-9 (Figure 2A). Median variations across manufacturers reached as high as 549% between AB and RO and the maximum variations reaching as high as 822% between AB and TH in May 2022. For other methods, the differences were lower as displayed in detail in Supplementary Table S3.

Similarly, these trends are even more evident in the relative collective medians of CA 19-9 when normalized to the overall median of the sample results (Figure 2B). Excluding the AB collective from the analysis substantially reduced the maximum manufacturer-specific differences to 222% when the DO collective, which had the highest value, is compared with the TH collective, which had the lowest value in May 2022. Notably, the AB collective exhibited the highest maximum normalized median difference of 7.61 and a median normalized median of 5.84, indicative of its substantial deviation from the overall median. Conversely, the medians for RO and TH for the normalized median were 0.93 and 0.96 respectively, and the TH collective displayed the lowest maximum normalized median difference of 1.29 (Supplementary Table S6).

The variation within individual collectives was, in fact, quiet low, with median CVs mostly below 10% (maximum CV 16%). This indicates a commendable level of assay precision (Figure 2C; Supplementary Table S5). Nevertheless, it should be noted that the SI collective sometimes displayed CVs as high as 36%. Dividing the SI collective into sub-collectives meant that the resulting subgroups, although still occasionally displaying CVs as high as 35% as in the case of the DG collective in January 2022 (Supplementary Figure S1B), had median CVs between 10% and 12% which is comparable to the other manufacturer collectives (Supplementary Table S7).

3.3 CA 125 EQA results

In the case of CA 125, either the AB or DO collective consistently reported the highest measured values for each EQA survey. A noteworthy change was observed in the performance of the BE collective, which consistently remained in the interquartile range of the overall box plot (grey box) before 2021, and then its values were only in the lower whisker range of all results (Figure 3A). Notable disparities in concentrations of CA 125 were observed across manufacturers, with median variations reaching as high as 99% between AB and BE and the maximum variations reaching as high as 151% between AB and BE in October 2021. For other methods, the differences were lower as displayed in detail in Supplementary Table S3.

In contrast to −20% to +20% before October 2021, the normalized median values of the BE collective from October 2021 onwards maintained a very consistent value for CA 125 measurements, with a bias of −30% in comparison to the overall median (Figure 3B). The AB collective had the highest normalized median value of 1.54, while the AX and RO collectives exhibited lower median values of 1.06 and 0.96 respectively (Supplementary Table S8).

Regarding method variability, the SI collective notably exhibited the highest scatter of results among the manufacturer collectives, with median CVs reaching 18% (maximum 25%). In contrast, the other collectives consistently maintained median CVs between 5% and 8% (maximum 20%; Figure 3C). As for the other CA markers studied in this paper, the high CVs of the SI collective went down once it was divided into its sub-collectives (Supplementary Figure S1C). The median CVs were 8% for BG, 6% for DG and 11% for SIE. Thus, they are more comparable to the median CVs of the other manufacturer collectives, which ranged from 5% to 8%, than to the overall SI collective with a median CV of 18% (Supplementary Table S9).

3.4 Evaluation of EQA results with respect to the current assessment limits

For CA 15-3, the AX collective displayed more variability, with the central 90% exceeding the limits in approximately half of the samples (Figure 4A). In contrast, the central 90% of values from the DO collective consistently adhered to the assessment limits for each sample (100% passing rate: ±24%), thereby demonstrating excellent performance (Figure 4B). The RO collective's results closely mirrored those of the DO collective, with only a minor deviation occurring twice when the central 90% were not able to pass the lower assessment limit (92% passing rate) (Figure 4C). The SI collective consistently exceeded the assessment limits in over half of the instances and the central 90% of SI passed the assessment limits only nine times (Figure 4D). Both collectives displayed fluctuations above and below the threshold. When evaluated separately, the three SI subtypes (BG, DG, SIE) had passing rates of 93%–96% (Table 1).

In the case of CA 19-9, the AB collective demonstrated exceptional consistency, with the central 90% not passing the lower assessment limit in just one instance (96% passing rate) (Figure 5A). On the other hand, the central 90% of values from the DO collective consistently remained within the assessment limits for each sample (100% passing rate: ±27%), demonstrating a robust performance (Figure 5B). Similar to CA 15-3, the RO collective delivered commendable results, with only one instance with the central 90% of laboratories being outside the upper assessment limit (96% passing rate) (Figure 5C). Notably, the SI collective consistently exceeded the assessment limit in every instance

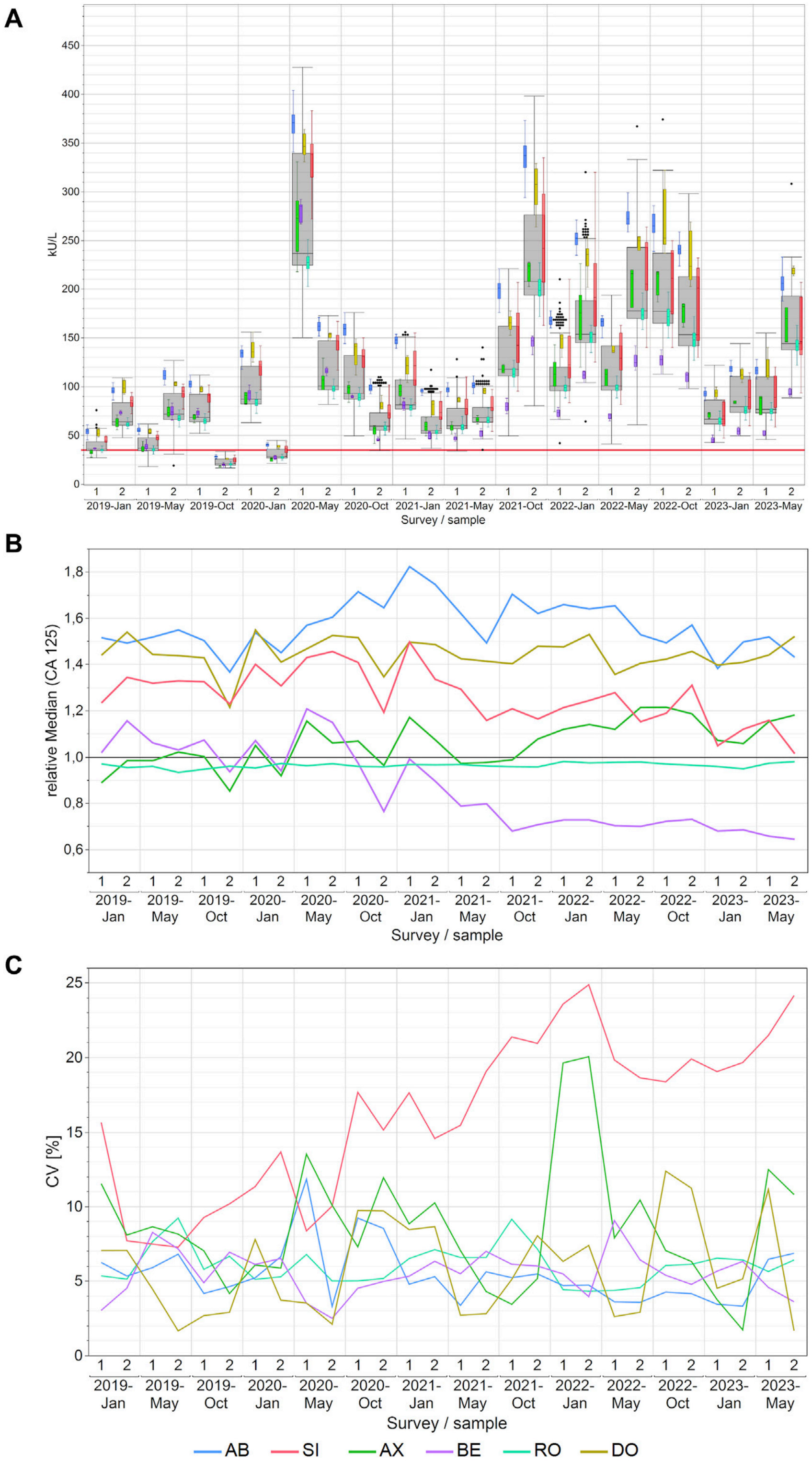


FIGURE 3
Manufacturer-dependent analysis of CA 125 EQA results, encompassing an all-results overview **(A)**, comparisons of manufacturer-dependent median differences relative to the overall median **(B)**, and evaluations of manufacturer-dependent CVs **(C)** between 2019 and 2023. Data are presented for two samples per survey. The gray boxes represent all results for the respective sample, while smaller, colored box plots overlay the total results (blue: AB, green: AX, cyan: RO, violet: BE, red: SI, ochre: DO). A red line marks the 35 kU/L cut-off value, and black dots denote outliers excluded from the colored boxes. The whiskers extend from the 1st quartile minus 1.5 times the interquartile range to the 3rd quartile plus 1.5 times the interquartile range.

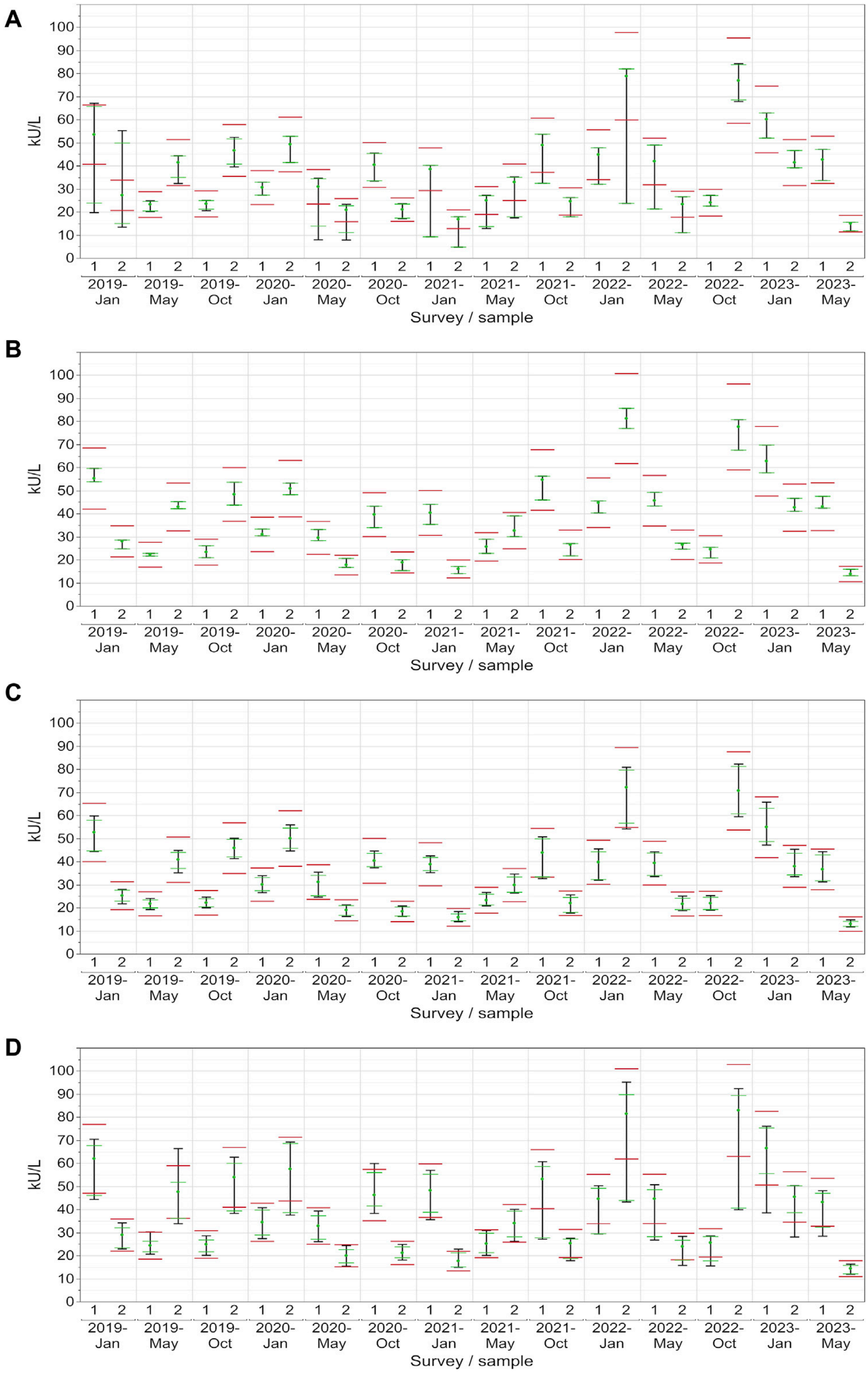


FIGURE 4
Manufacturer-specific evaluation of EQA results for CA 15-3 with respect to the current assessment limits for the AX **(A)**, DO **(B)**, RO **(C)** and SI **(D)** collectives. The green dot represents the median of all results within each respective collective and EQA survey. Assessment limits of ±24% are highlighted with red lines, while green lines indicate the median for 80% of the results, and a black line signifies the median for 90% of the results.

except once in October 2020 (Figure 5D). When evaluated separately, the three SI subtypes (BG, DG, SIE) had passing rates of 68%–82% (Table 1).

When looking at CA 125, the AB collective also performed comparably well, with only a single instance where the central 90% slightly did not pass the upper assessment limit (96% passing rate)

TABLE 1 Evaluation of EQA results with respect to the current assessment limits and stricter, uniform assessment limits of 24%.

Manufacturer	CA 15-3 (±24%)	CA 19-9 (±27%)	CA 19-9 (±24%)	CA 125 (±36%)	CA 125 (±24%)
	passing (5%–95%) [n/%]	passing (5%–95%) [n/%]	passing (5%–95%) [n/%]	passing (5%–95%) [n/%]	passing (5%–95%) [n/%]
AB	24/85.7	27/96.4	26/92.9	27/96.4	25/89.3
AX	14/50.0	24/85.7	24/85.7	26/92.9	22/78.6
BE	22/78.6	26/92.9	23/82.1	28/100.0	28/100.0
DO	28/100.0	28/100.0	28/100	28/100.0	26/92.9
RO	26/92.9	27/96.4	24/85.7	28/100.0	28/100.0
SI	9/32.1	1/3.6	1/3.6	18/64.3	8/28.6
SI—BG	26/92.9	23/82.1	18/64.3	25/89.3	18/64.3
SI—DG	26/92.9	21/75	18/64.3	28/100.0	26/92.9
SI—SIE	26/96.4	19/67.9	19/67.9	25/89.3	19/67.9
TH		24/85.7	22/78.6		

(Figure 6A). Both the RO and DO collectives consistently maintained all values within the assessment limits for each sample (100% passing rate: ±36%), which reflects a strong performance (Figures 6B, C). Even the central 90% of the more variable SI collective exceeded the assessment limit on only 10 out of 28 occasions (Figure 6D). When evaluated separately, the three SI subtypes (BG, DG, SIE) had passing rates of 89%–100% (Table 1).

When the assessment limits were adjusted so that each tumor marker had the 24% passing limit as stipulated for CA 15-3 by the RiliBÄK guidelines (Bundesärztekammer, 2023), the central 90% of most collectives would still pass on many occasions (Table 1; Supplementary Figures S5, S6) with passing rates of 79%–100% for most manufacturers for CA 19-9 (only the 3 SI subtypes remained below 70%) and 79%–100% for most manufacturers for CA 125 (only 2 of the SI subtypes remained below 70%).

4 Discussion

The utilization of EQA material for comparative analysis provides a standardized framework for evaluating laboratory performance across different assays. While some EQA institutions in other countries use patient samples, similar phenomena and variations are observed for both materials (van Rossum et al., 2024). This study undertakes a thorough re-evaluation of recent EQA data spanning from 2019 to 2023 for the biomarkers CA 15-3, CA 19-9 and CA 125 and highlights notable variations in the performance of tumor marker assays.

The high variability across manufacturers for CA 15-3 was also reported by Slev et al., who performed a comparative analysis of seven automated CA assays and found BE consistently yielding lower results than the SI sub-collective BG (Slev et al., 2006). Similarly, clinical studies have demonstrated considerable method dependent differences for CA 15-3 (Molina et al., 2008), CA 19-9 (Stieber et al., 2008b) and CA 125 (Holdenrieder et al., 2008).

Potential causes of these manufacturer-related differences include the utilization of distinct monoclonal antibodies across assays with different binding sites and affinities due to variable antigen-binding sites, as well as antigen modifications such as glycosylation and different assay formulations (Price et al., 1998; Reinauer and Wood, 2005; Partyka et al., 2012; Zeng et al., 2012; Wojtalewicz et al., 2023). Fortunately, high intra-manufacturer consistency with CVs below 16% was found for all methods studied. This is particularly beneficial when the same methods are applied for monitoring individual patients over time. However, any transition to another method should be carefully managed with double measurements using both methods to minimize disruptions in patient care and ensure continuity in result interpretation. Notably, the low CVs observed here align with similar trends seen in previous marker analyses, such as AFP and CEA, where even lower CVs were observed (Wojtalewicz et al., 2023). Given that certified reference materials (CRM) for AFP and CEA already exist, it is expected that further improvements of CVs for CA 15-3, CA 19-9 and CA 125 will occur once CRMs for these markers are developed (Sturgeon, 2016).

It is important to highlight that the consistent CVs, the high passing rates of the EQA schemes and the considerable differences between the methods remained stable over the studied time interval. A comparison between the present study and an earlier one conducted in 2005 (Reinauer and Wood, 2005) revealed some changes over the past 2 decades. The maximum differences observed were 162% for CA 125, 44% for CA 15%–3% and 195% for CA 19-9. Therefore, manufacturers are urgently called upon to improve the standardization and harmonization of their methods and regulative bodies are encouraged to provide CRMs as a basis for more accurate alignment.

Furthermore, it is imperative that manufacturers conduct clinical performance studies for their tumor marker assays. These studies are essential not only to establish method-specific decision limits for reference intervals in healthy individuals, but also to evaluate criteria for distinguishing

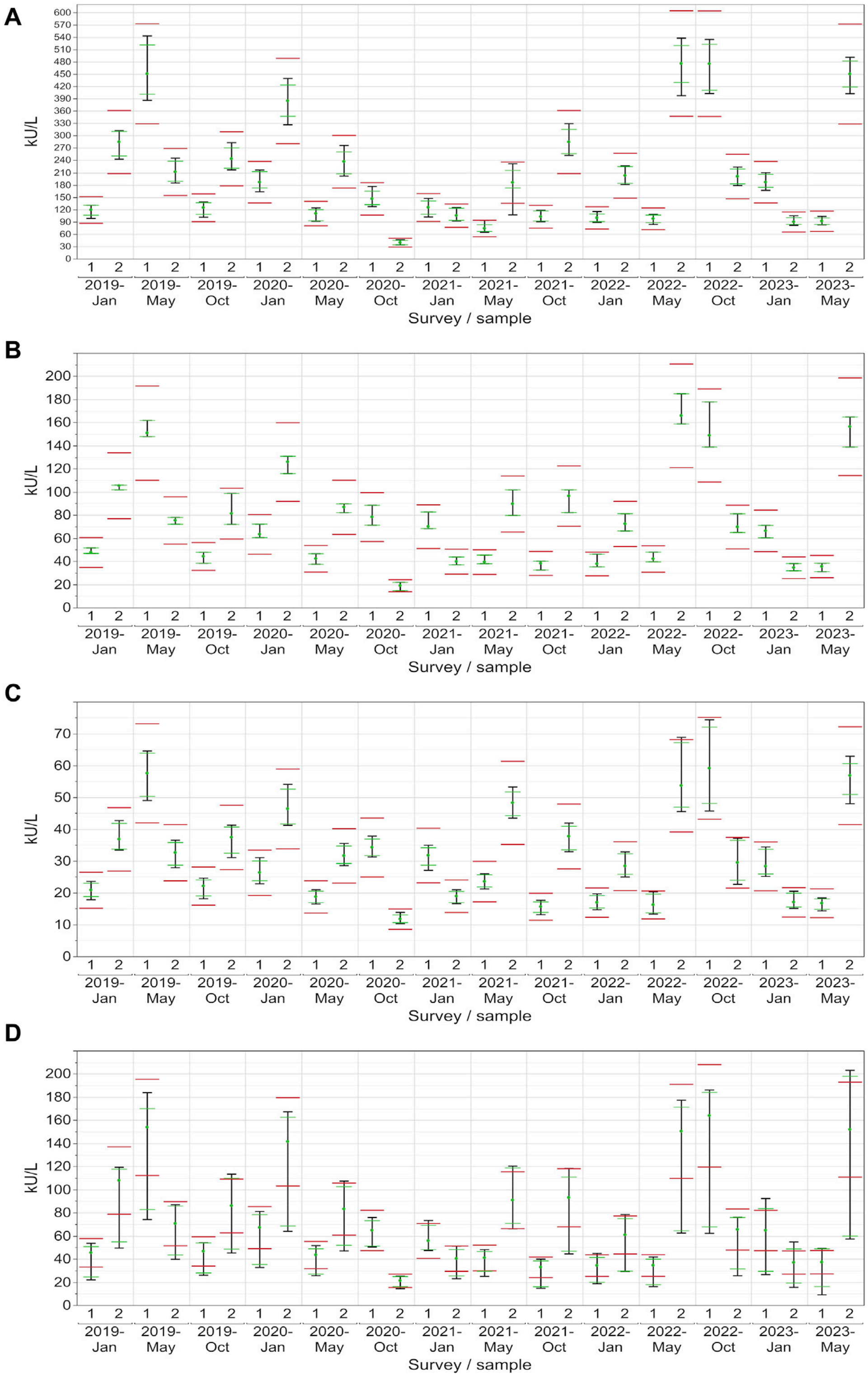


FIGURE 5
Manufacturer-specific evaluation of EQA results for CA 19-9 with respect to the current assessment limits for the AB **(A)**, DO **(B)**, RO **(C)** and SI **(D)** collectives. The green dot represents the median of all results within each respective collective and EQA survey. Assessment limits of ±27% are highlighted with red lines, while green lines indicate the median for 80% of the results, and a black line signifies the median for 90% of the results.

between malignant and benign conditions. Additionally, it is crucial to develop criteria for estimating prognosis at different stages of disease and to assess relative increases or decreases in individual patients to measure therapeutic efficacy. This is highly important, as clinical decision criteria will differ for each indication of tumor marker application in cancer

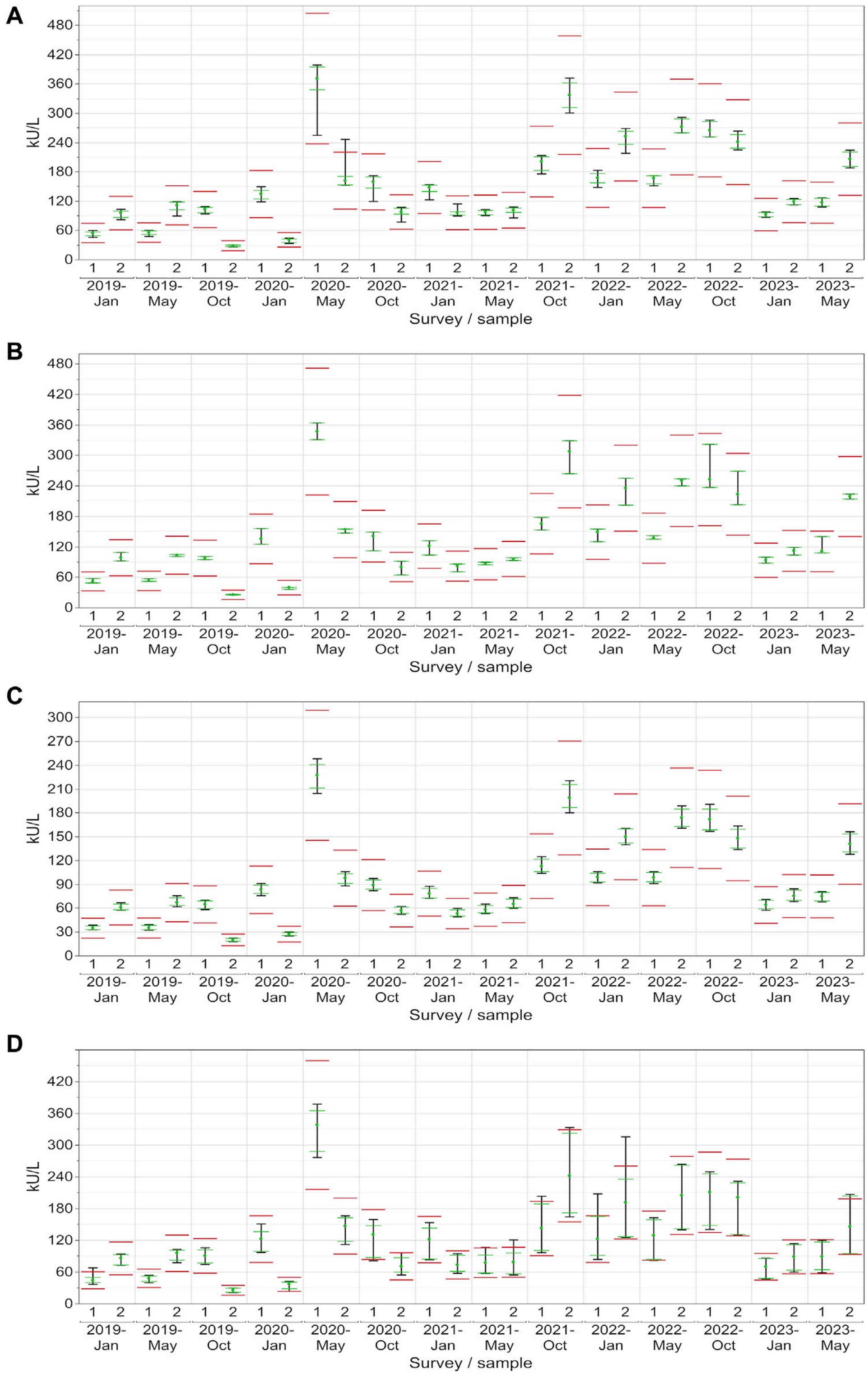


FIGURE 6
Manufacturer-specific evaluation of EQA results for CA 125 with respect to the current assessment limits for the AB **(A)**, DO **(B)**, RO **(C)** and SI **(D)** collectives. The green dot represents the median of all results within each respective collective and EQA survey. Assessment limits of ±36% are highlighted with red lines, while green lines indicate the median for 80% of the results, and a black line signifies the median for 90% of the results.

patients. Given the considerable variability among individual methods, such studies will enhance the clinical relevance of the assays and optimize their use in patient care.

When differences between methods were related to a normalized median of all methods, a certain bias has to be taken into account, as the RO collective was overrepresented in the whole cohort.

Divergent trends in relative medians across individual groups may be attributed to factors such as interfering substances, matrix effects and molecular heterogeneity, particularly for CA 19-9 (Denis et al., 2019; Mahadevarao Premnath and Zubair, 2024). Higher CVs in individual methods can be attributed to interfering substances (Sturgeon and Viljoen, 2011), the simple fact of low participant numbers and variances in assay lot calibration. As reported by Kim et al., the lot effect can result in variances up to 14.3% for CA 19-9 (Kim et al., 2012).

Consequently, the commutability of EQA materials with patient samples is crucial. EQA samples were produced using a human serum-like matrix spiked with the respective tumor antigens from cell cultures. Importantly, the observed manufacturer-specific variations are not necessarily attributable to the spiked material, as similar differences in methods were also observed in plasma samples (van Rossum et al., 2024), with consistently higher concentration of CA 19-9 for AB compared to other manufacturers. Nevertheless, a commutability study with a direct comparison of artificial and patient material is still pending.

Currently, only the EQA acceptance criteria of ±24% for CA 15-3 are defined in the German Medical Association's RiliBÄK guideline, while criteria for CA 19-9 and CA 125 are not specified (Bundesärztekammer, 2023). Historically, higher acceptance ranges of ±27% for CA 19–9 and ±36% for CA 125 have been defined. These criteria have allowed almost all participants to regularly pass the EQA schemes. However, such broad ranges mean that changes up to 72% for CA 125 might not be interpreted as genuine disease-related changes in individual patients, given the high potential for analytical variability–even when using the same method. Therefore, more stringent limits would be beneficial to enable the clinical interpretation of already smaller dynamic tumor marker changes in individual patients. This approach could help to prevent misdiagnosis and unnecessary invasive tests, as has been discussed in the context of HbA1c measurements (Heinemann et al., 2018).

However, if the limit of ±24% was applied to all three markers, the majority of participants would still pass the EQAs. In contrary, the low variability within methods suggests that even more stringent limits could be feasible. Narrowing the acceptance criteria would improve the quality and reliability of clinical decision-making when interpreting individual tumor marker dynamics. This would be especially relevant for monitoring therapy progress in cancer patients or for disease monitoring after tumor removal. With the new acceptance criteria, changes of 50% could be interpreted reliably. However, this necessitates maintaining consistent methods over longitudinal courses, clearly indicating these methods in laboratory reports and ensuring their inclusion in electronic reports together with the measured values. Furthermore, this information should be incorporated into the newly introduced electronic patient records on a nationwide basis in Germany.

In addition to these measures, manufacturers are encouraged to enhance the standardization and harmonization of tumor marker assays. This includes minimizing manufacturer-specific differences, optimizing assay performance, and conducting clinical studies. Continued collaboration between manufacturers, regulatory agencies, professional organizations, and clinical laboratories is crucial for advancing the field of tumor marker testing and improving the quality of patient care (Aarsand and Sandberg, 2014; Tate et al., 2014; Ceriotti, 2016; Plebani, 2016).

Laboratories within the public health network often encounter challenges during procurement processes, where price considerations may overshadow concerns regarding assay quality and performance. It is crucial to emphasize that tumor marker diagnostics are only valuable if the assays used meet the highest quality standards which should outweigh economic considerations. The results of this longitudinal EQA analysis comparing different methods and manufacturers provide compelling arguments for selecting appropriate assays. These findings may also encourage manufacturers to prioritize assay performance and reliability when developing and calibrating tumor marker assays, thereby enhancing the quality of oncological diagnostics in public health laboratories.

5 Conclusion

The present study provides a large set of longitudinal data from EQA schemes for tumor markers CA 15-3, CA 19-9 and CA 125 assessed by different methods and manufacturers. While intra-manufacturer variability was acceptable, inter-manufacturer variability was quite high, which has severe consequences for application of tumor markers in patient care. Therefore, better standardization and harmonization are urgently needed. The development of CRMs and continuous guidance by regulatory bodies will support this process, necessitating close collaboration between manufacturers, regulatory agencies, professional scientific organizations, and clinical laboratories.

Beyond analytical and preanalytical validation, comprehensive clinical studies on the performance of tumor marker tests as well as the definition of meaningful clinical decision criteria for various indications throughout the course of cancer are essential. Improved and internationally aligned acceptance criteria for passing EQA schemes will enable a qualified and sensitive interpretation of longitudinal marker changes in individual cancer patients. These quality indicators are fundamental and should always take precedence over economic consideration. Only through the collaborative efforts of all stakeholders striving for higher quality standards can diagnostic guidance for cancer patients be improved.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical approval was not required for the studies involving humans because only commercially available established serum pool samples were used. The studies were conducted in accordance with the local legislation and institutional requirements. The human samples used in this study were acquired from a commercial quality control sample provider. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

MK: Writing–original draft, Writing–review and editing, Conceptualization, Formal Analysis, Investigation, Methodology, Validation, Visualization. NW: Writing–original draft, Writing–review and editing, Conceptualization, Investigation, Methodology, Validation. AK-S: Writing–original draft, Writing–review and editing, Investigation, Validation. LV: Writing–original draft, Writing–review and editing, Investigation, Methodology, Validation. AS: Writing–original draft, Writing–review and editing. IS: Writing–original draft, Writing–review and editing, Conceptualization, Resources, Supervision. SH: Writing–original draft, Writing–review and editing, Conceptualization, Investigation, Resources, Supervision, Validation.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2024.1401619/full#supplementary-material

References

Aarsand, A. K., and Sandberg, S. (2014). How to achieve harmonisation of laboratory testing -The complete picture. *Clin. Chim. Acta* 432, 8–14. doi:10.1016/j.cca.2013.12.005

Bray, F., Laversanne, M., Weiderpass, E., and Soerjomataram, I. (2021). The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* 127 (16), 3029–3030. doi:10.1002/cncr.33587

Bundesärztekammer (2023). Richtlinie der Bundesärztekammer zur Qualitätssicherung laboratoriumsmedizinischer Untersuchungen. *Dtsch. Ärzteblatt Jg.* 120, 21–22. doi:10.3238/arztebl.2023.rili_baek_QS_Labor

Ceriotti, F. (2016). Harmonization initiatives in europe. *EJIFCC* 27 (1), 23–29.

Charkhchi, P., Cybulski, C., Gronwald, J., Wong, F. O., Narod, S. A., and Akbari, M. R. (2020). CA125 and ovarian cancer: a comprehensive review. *Cancers (Basel)* 12 (12), 3730. doi:10.3390/cancers12123730

DelVillano, B. C., and Zurawski, V. R., Jr. (1983). The carbohydrate antigenic determinant 19-9 (CA 19-9): a monoclonal antibody defined tumor marker. *Lab. Res. Methods Biol. Med.* 8, 269–282.

Denis, J. A., Mazzola, A., Nguyen, G., Lacorte, J. M., Brochet, C., Larsen, A. K., et al. (2019). Transient increase of CA 19-9 serum concentrations in a liver transplant recipient with cystic fibrosis and hepatic abscess: a case report and brief literature review. *Clin. Biochem.* 64, 53–56. doi:10.1016/j.clinbiochem.2018.10.009

Duffy, M. J., Evoy, D., and McDermott, E. W. (2010). CA 15-3: uses and limitation as a biomarker for breast cancer. *Clin. Chim. Acta* 411 (23-24), 1869–1874. doi:10.1016/j.cca.2010.08.039

Duffy, M. J., Shering, S., Sherry, F., McDermott, E., and O'Higgins, N. (2000). CA 15-3: a prognostic marker in breast cancer. *Int. J. Biol. Markers* 15 (4), 330–333. doi:10.1177/172460080001500410

Escudero, J. M., Auge, J. M., Filella, X., Torne, A., Pahisa, J., and Molina, R. (2011). Comparison of serum human epididymis protein 4 with cancer antigen 125 as a tumor marker in patients with malignant and nonmalignant diseases. *Clin. Chem.* 57 (11), 1534–1544. doi:10.1373/clinchem.2010.157073

Filella, X., Rodriguez-Garcia, M., and Fernandez-Galan, E. (2023). Clinical usefulness of circulating tumor markers. *Clin. Chem. Lab. Med.* 61 (5), 895–905. doi:10.1515/cclm-2022-1090

Gang, Y., Adachi, I., Ohkura, H., Yamamoto, H., Mizuguchi, Y., and Abe, K. (1985). CA 15-3 is present as a novel tumor marker in the sera of patients with breast cancer and other malignancies. *Gan Kagaku Ryoho* 12 (12), 2379–2386.

Ghosh, I., Bhattacharjee, D., Das, A. K., Chakrabarti, G., Dasgupta, A., and Dey, S. K. (2013). Diagnostic role of tumour markers CEA, CA15-3, CA19-9 and CA125 in lung cancer. *Indian J. Clin. Biochem.* 28 (1), 24–29. doi:10.1007/s12291-012-0257-0

Heinemann, L., Kaiser, P., Freckmann, G., Grote-Koska, D., Kerner, W., Landgraf, R., et al. (2018). Higher HbA1c measurement quality standards are needed for follow-up and diagnosis: experience and analyses from Germany. *Horm. Metab. Res.* 50 (10), 728–734. doi:10.1055/a-0721-2273

Holdenrieder, S., Molina, R., Gion, M., Gressner, A., Troalen, F., Auge, J. M., et al. (2008). Alternative antibody for the detection of CA125 antigen: a European multicenter study for the evaluation of the analytical and clinical performance of the Access OV Monitor assay on the UniCel Dxl 800 Immunoassay System. *Clin. Chem. Lab. Med.* 46 (5), 588–599. doi:10.1515/CCLM.2008.125

ISO13528 (2022). Statistical methods for use in proficiency testing by interlaboratory comparisons. *Int. Organ. Stand. (ISO)*. Available at: https://www.iso.org/standard/78879.html

Kim, H. S., Kang, H. J., Whang, D. H., Lee, S. G., Park, M. J., Park, J. Y., et al. (2012). Analysis of reagent lot-to-lot comparability tests in five immunoassay items. *Ann. Clin. Lab. Sci.* 42 (2), 165–173.

Kim, S., Park, B. K., Seo, J. H., Choi, J., Choi, J. W., Lee, C. K., et al. (2020). Carbohydrate antigen 19-9 elevation without evidence of malignant or pancreatobiliary diseases. *Sci. Rep.* 10 (1), 8820. doi:10.1038/s41598-020-65720-8

La'ulu, S. L., and Roberts, W. L. (2007). Performance characteristics of five automated CA 19-9 assays. *Am. J. Clin. Pathol.* 127 (3), 436–440. doi:10.1309/H52VET3M6P7GYWG1

Lee, T., Teng, T. Z. J., and Shelat, V. G. (2020). Carbohydrate antigen 19-9 - tumor marker: past, present, and future. *World J. Gastrointest. Surg.* 12 (12), 468–490. doi:10.4240/wjgs.v12.i12.468

Mahadevarao Premnath, S., and Zubair, M. (2024). Laboratory evaluation of tumor biomarkers. StatPearls. Treasure island (FL): StatPearls. Available at: https://www.ncbi.nlm.nih.gov/books/NBK597378/ (Accessed February 1, 2024).

Molina, R., Gion, M., Gressner, A., Troalen, F., Auge, J. M., Holdenrieder, S., et al. (2008). Alternative antibody for the detection of CA15-3 antigen: a European multicenter study for the evaluation of the analytical and clinical performance of the Access BR Monitor assay on the UniCel Dxl 800 Immunoassay System. *Clin. Chem. Lab. Med.* 46 (5), 612–622. doi:10.1515/CCLM.2008.133

Mongia, S. K., Rawlins, M. L., Owen, W. E., and Roberts, W. L. (2006). Performance characteristics of seven automated CA 125 assays. *Am. J. Clin. Pathol.* 125 (6), 921–927. doi:10.1309/NBA3-12W0-LANR-XYH9

Moore, R. G., McMeekin, D. S., Brown, A. K., DiSilvestro, P., Miller, M. C., Allard, W. J., et al. (2009). A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass. *Gynecol. Oncol.* 112 (1), 40–46. doi:10.1016/j.ygyno.2008.08.031

Partyka, K., Maupin, K. A., Brand, R. E., and Haab, B. B. (2012). Diverse monoclonal antibodies against the CA 19-9 antigen show variation in binding specificity with consequences for clinical interpretation. *Proteomics* 12 (13), 2212–2220. doi:10.1002/pmic.201100676

Passerini, R., Riggio, D., Salvatici, M., Zorzino, L., Radice, D., and Sandri, M. T. (2007). Interchangeability of measurements of CA 19-9 in serum with four frequently used assays: an update. *Clin. Chem. Lab. Med.* 45 (1), 100–104. doi:10.1515/CCLM.2007.003

Plebani, M. (2016). Harmonization of clinical laboratory information - current and future strategies. *EJIFCC* 27 (1), 15–22.

Price, M. R., Rye, P. D., Petrakou, E., Murray, A., Brady, K., Imai, S., et al. (1998). Summary report on the ISOBM TD-4 Workshop: analysis of 56 monoclonal antibodies against the MUC1 mucin. San Diego, Calif., November 17-23, 1996. *Tumour Biol.* 19 (Suppl. 1), 1–20. doi:10.1159/000056500

Reinauer, H., and Wood, W. G. (2005). External quality assessment of tumour marker analysis: state of the art and consequences for estimating diagnostic sensitivity and specificity. *Ger. Med. Sci.* 3, Doc02.

Serdarevic, N. (2018). The comparison between different immunoassays for serum carbohydrate antigen (CA 19-9) concentration measurement. *Acta Inf. Med.* 26 (4), 235–239. doi:10.5455/aim.2018.26.235-239

Slev, P. R., Rawlins, M. L., and Roberts, W. L. (2006). Performance characteristics of seven automated CA 15-3 assays. *Am. J. Clin. Pathol.* 125 (5), 752–757. doi:10.1309/G6X6-PR75-26FA-KV0E

Stieber, P., and Heinemann, V. (2008a). Sinnvoller Einsatz von Tumormarkern/Sensible use of tumor markers. *J. Laboratory Med.* 32 (5), 339–360. doi:10.1515/jlm.2008.015

Stieber, P., Molina, R., Gion, M., Gressner, A., Troalen, F., Holdenrieder, S., et al. (2008b). Alternative antibody for the detection of CA19-9 antigen: a European multicenter study for the evaluation of the analytical and clinical performance of the Access GI Monitor assay on the UniCel Dxl 800 Immunoassay System. *Clin. Chem. Lab. Med.* 46 (5), 600–611. doi:10.1515/CCLM.2008.126

Sturgeon, C. (2016). Standardization of tumor markers - priorities identified through external quality assessment. *Scand. J. Clin. Lab. Investig. Suppl.* 245, S94–S99. doi:10.1080/00365513.2016.1210334

Sturgeon, C. M., and Viljoen, A. (2011). Analytical error and interference in immunoassay: minimizing risk. *Ann. Clin. Biochem.* 48 (Pt 5), 418–432. doi:10.1258/acb.2011.011073

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Tate, J. R., Johnson, R., Barth, J. H., and Panteghini, M. (2014). Harmonization of laboratory testing - a global activity. *Clin. Chim. Acta* 432, 1–3. doi:10.1016/j.cca.2014.02.006

Trape, J., Filella, X., Alsina-Donadeu, M., Juan-Pereira, L., Bosch-Ferrer, A., Rigo-Bonnin, R., et al. (2011). Increased plasma concentrations of tumour markers in the absence of neoplasia. *Clin. Chem. Lab. Med.* 49 (10), 1605–1620. doi:10.1515/CCLM.2011.694

van Rossum, H. H., Holdenrieder, S., Ballieux, B. E. P. B., Badrick, T. C., Yun, Y.-M., Zhang, C., et al. (2024). Investigating the current harmonization status of tumor markers using global external quality assessment programs: a feasibility study. *Clin. Chem.* 70, 669–679. doi:10.1093/clinchem/hvae005

Wojtalewicz, N., Vierbaum, L., Kaufmann, A., Schellenberg, I., and Holdenrieder, S. (2023). Longitudinal evaluation of AFP and CEA external proficiency testing reveals need for method harmonization. *Diagn. (Basel)* 13 (12), 2019. doi:10.3390/diagnostics13122019

Zeng, X., Shen, Z., and Mernaugh, R. (2012). Recombinant antibodies and their use in biosensors. *Anal. Bioanal. Chem.* 402 (10), 3027–3038. doi:10.1007/s00216-011-5569-z

TYPE Original Research
PUBLISHED 21 June 2024
DOI 10.3389/fmolb.2024.1390079

OPEN ACCESS

EDITED BY
Matteo Becatti,
University of Firenze, Italy

REVIEWED BY
Sidra Islam,
Cleveland Clinic, United States
Marijana Miler,
Sestre milosrdnice University Hospital Center,
Croatia

*CORRESPONDENCE
Luisa Toll,
✉ luisa.toll@hs-furtwangen.de

RECEIVED 22 February 2024
ACCEPTED 10 May 2024
PUBLISHED 21 June 2024

CITATION
Toll L, Weiss N, Vierbaum L, Schellenberg I, Thevis M and Wenzel F (2024), Longitudinal evaluation of laboratory results and method precision in worldwide erythropoietin external quality assessments.
Front. Mol. Biosci. 11:1390079.
doi: 10.3389/fmolb.2024.1390079

Longitudinal evaluation of laboratory results and method precision in worldwide erythropoietin external quality assessments

Luisa Toll[1,2]*, Nathalie Weiss[3], Laura Vierbaum[3], Ingo Schellenberg[3,4], Mario Thevis[2] and Folker Wenzel[1,3]

[1]Faculty of Medical and Life Sciences, Furtwangen University, Villingen-Schwenningen, Germany, [2]Institute of Biochemistry/ Center for Preventive Doping Research, German Sport University Cologne, Cologne, Germany, [3]INSTAND e.V., Society for Promoting Quality Assurance in Medical Laboratories, Düsseldorf, Germany, [4]Institute of Bioanalytical Sciences (IBAS), Center of Life Sciences, Anhalt University of Applied Sciences, Bernburg, Germany

Introduction: This study presents a longitudinal analysis of external quality assessment (EQA) results for erythropoietin (EPO) determinations conducted between 2017 and 2022 with a continuously increasing number of participating laboratories. The aim of this work was to evaluate participant performance and methodological aspects.

Methods: In each of the eleven EQA surveys, a blinded sample set of lyophilized human serum containing one sample with lower EPO concentrations (L) and one with higher EPO concentrations (H) was sent to the participating laboratories.

Results: A total of 1,256 measurements were included. The median (interquartile range) fraction of participants not meeting the criteria of acceptance set at 20% around the robust mean of the respective survey was 9.5% (6.1%–10.7%) (sample L) and 9.1% (5.8%–11.8%) (sample H) but lacked a clear trend in the observed period. Some surveys exhibited unusually high interlaboratory variation, suggesting interfering components in the EQA samples. Different immunological methods and reagent manufacturers also showed variability in measurement outcomes to some extent.

Conclusion: These findings highlight the need for continuous quality assessment in EPO measurements to ensure patient safety and identify areas for further research and investigation.

KEYWORDS

external quality assessment, proficiency testing, erythropoietin, method precision, immunoassay

1 Introduction

The quantitative determination of erythropoietin (EPO) in blood is mainly performed using immunoassays. By measuring serum EPO levels, useful information can be obtained on various pathogenic changes. The resulting therapeutic algorithms can guide treatment. Chronic kidney disease, as well as systemic inflammation and malignancies, can lead to a decrease in EPO biosynthesis and, therefore, to low EPO levels in the blood (Jelkmann, 2011; Portolés et al., 2021). Higher concentrations can be measured for secondary erythrocytosis, mostly caused by hypoxemia. In addition, non-renal anemia results in a higher renal EPO production and an exponential increase in serum levels (Artunc and Risler, 2007; Jelkmann, 2011; Bunn, 2013). In combination with other parameters, EPO also serves as a marker for possible myeloproliferative diseases (Michiels et al., 2007). Endogenous EPO levels should also be determined before injecting erythropoiesis-stimulating agents to treat, for example, myelodysplasia (Fried, 2009).

An adequate measurement quality is essential for ensuring patient safety, and a formal proof of the analytical competence to measure certain parameters—namely, accreditation—is mandatory or at least recommended in most countries (Zima, 2017). Adequate treatment and patient safety require reliable test results at a consistently high standard (Laudus et al., 2022). Clinicians and especially patients expect precise test results from diagnostic testing during treatment monitoring, regardless of the laboratory performing the tests (De La Salle et al., 2017). External quality assessment (EQA) is used to independently evaluate, continuously monitor, and compare laboratory performance, and frequent participation in EQA programs is mandatory for accredited medical laboratories (Favaloro et al., 2018; Sciacovelli et al., 2018; DIN EN ISO 15189:2023-03, 2023). It is a helpful tool for accessing the current *status quo* and can help to identify areas in need of improvement (Laudus et al., 2022). Additionally, EQA can assess the precision of the methodology used by the laboratories (Favaloro et al., 2018).

Marsden et al. emphasized the need for the establishment of an EPO EQA scheme in 2006 after they found fluctuations of 2.9–200 IU/L in a sample distribution program involving six laboratories (Marsden et al., 2006). INSTAND e.V. is an independent scientific medical society and accredited organization located in Germany supporting quality assurance in medical laboratories by performing EQA in laboratory medicine. INSTAND introduced its first worldwide EQA for EPO measurements in 2017. Since then, it has been performed twice a year, and the certificate is valid for 12 months.

In order to observe developments in the general measurement quality of medical laboratories and their methodology for EPO measurement, an established EQA scheme with a certain number of participants and EQA runs is required. This study is the first to show a longitudinal analysis of the results of the INSTAND EPO EQA from 2017 to 2022 with participating laboratories from all over the world. The study also aims to summarize the results of all runs of this EQA and to present the development of the EPO EQA since its introduction.

2 Materials and methods

2.1 EPO EQA procedure

A total of eleven surveys of the EPO EQA were performed twice per year (surveys S1/S2) between 2017 and 2022, which involved an increasing number of participants from all over the world. For each survey, every participating laboratory was asked to analyze two blinded lyophilized human serum samples containing different EPO concentrations. In this work, the sample with the lower concentration is always referred to as sample L, and the one with the higher concentration is sample H. In some cases, specimens were enriched with recombinant EPO by the sample manufacturer. Due to an unexpectedly high number of participants in 2020-S1 and 2020-S2, participants were divided into two subsets (2020-S1a and 2020-S1b and 2020-S2a and 2020-S2b), and each received a different sample set. The lyophilized EQA samples had to be reconstituted with 1 mL of distilled water for 30 min at room temperature and then analyzed like a normal patient sample.

Laboratories reported their results and information about the assay they used to INSTAND via the RV-Online platform (https://rv-online.instandev.de). Between 2017 and 2022, the EQA criteria of acceptance (CoA) for EPO were set to a 20% deviation from the robust mean calculated using Algorithm A (ISO13528:2015, 2020). Laboratories that reported measurements outside the CoA would not pass the quality assessment. The German Medical Association has not yet defined a maximum permissible relative deviation in EQA schemes for EPO. Therefore, the CoA used for the evaluation of the INSTAND EPO EQA is based on the mean value of the permissible relative deviations recommended in the guideline of the German Medical Association for EQA schemes for other quantitative parameters in clinical chemistry (Bundesärztekammer, 2022).

2.2 Data analysis

Microsoft Excel (Version 16.56, Microsoft Corporation, Redmond, WA, USA) was used for data management. The statistical analysis and visualization of the results were performed using R Studio (Version 4.1.1 (2021-08–10), Rstudio PBC, Boston, MA, USA). Figures were created using the R-package ggplot2 (Wickham, 2016). The whiskers in the created boxplots span 1.5 times the interquartile range (IQR) above and below the box, capturing the middle 50% of the data. The dots mark outliers, which are defined as observations that exceed 1.5 times the IQR from either edge of the box.

The mean absolute deviation (MAD) to median ratio was calculated to evaluate the interlaboratory variation. Data distribution depending on the immunological methods used by the laboratories was analyzed. Methods used by the participating laboratories were enzyme-linked immunosorbent assay (ELISA), chemiluminescence immunoassay (CLIA), or luminescent enzyme immunoassay (LEIA). Reagent manufacturers' dependent data distributions were analyzed. The manufacturers were Beckman Coulter, Inc. (BE), Siemens (DPC-Biermann; DG), and IBL International GmbH (IB). Missing information on test method and reagent manufacturer, as well as manufacturer collectives with $n < 14$, were grouped as "other" due to lack of statistical validity.

Nine measurements each for sample L and sample H were excluded from the dataset due to suspected sample mix-ups or data submission errors and were not included in later calculations (Supplementary Table S2).

3 Results

Overall, 1,256 measurements were evaluated. The first EQA survey conducted in 2017 had ten participating laboratories. In

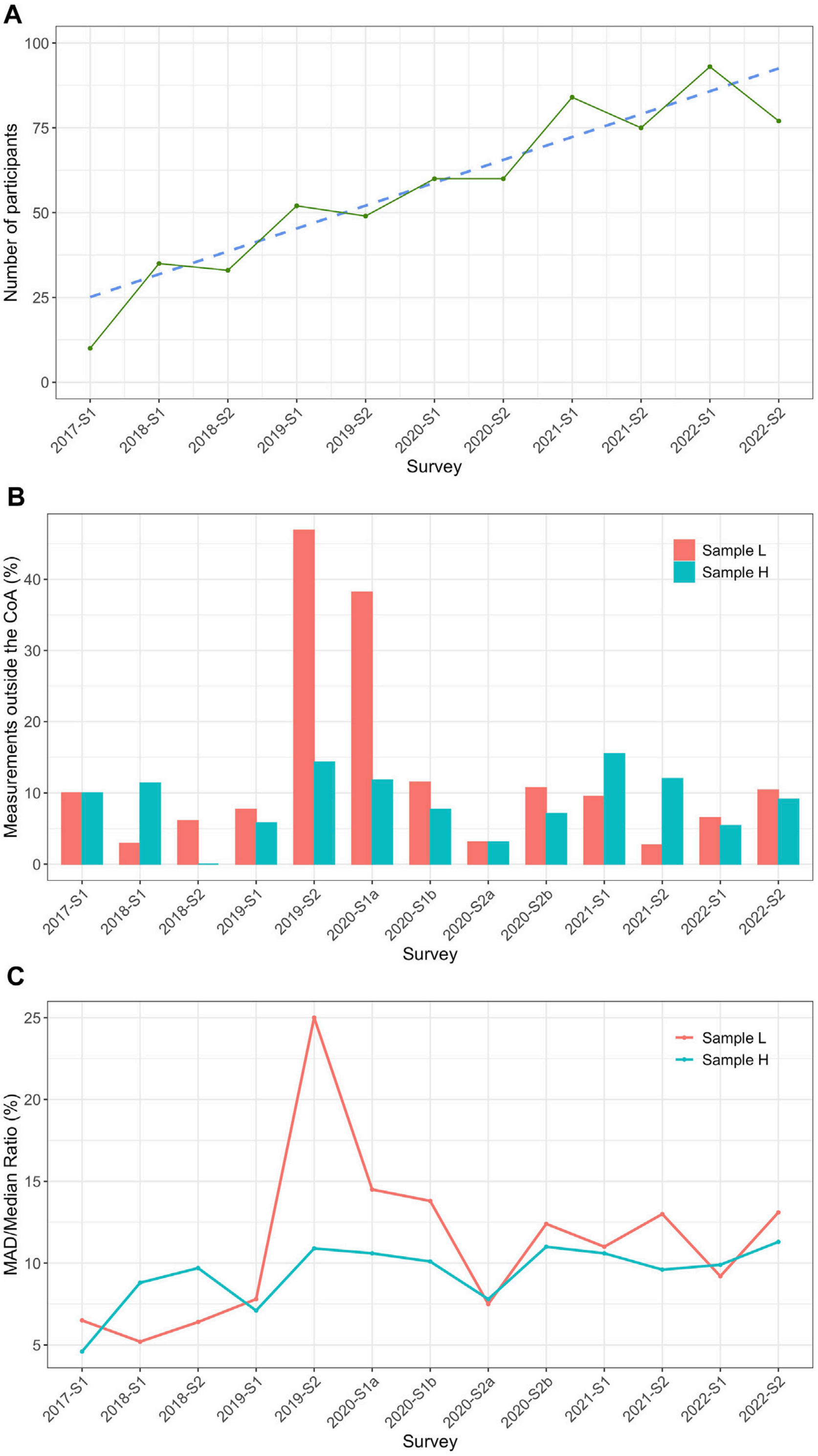


FIGURE 1
General outcome/information of the INSTAND EPO EQA from 2017 to 2022. **(A)** Number of laboratories participating between 2017 and 2022 (green) and a corresponding trend line (blue) starting with one survey (S) in 2017 (2017-S1) and continuing with two runs per year (S1/S2) until 2022. **(B)** The percentage of measurements outside the criteria of acceptance (CoA; %) calculated for each survey for sample L (red) and sample H (turquoise). The CoA was defined as ± 20% around the robust mean for the individual surveys shown. **(C)** Mean absolute deviation (MAD)/median ratio (%) for sample L (red) and sample H (turquoise) for every survey.

TABLE 1 Robust mean values (IU/L) calculated by Algorithm A (ISO13528:2015, 2020) and measurements outside of the criterion of acceptance (CoA) at ± 20% around the robust mean for each of the eleven surveys (S) from 2017 to 2022 for sample L (L) and sample H (H).

Survey	Robust mean (Algorithm A; IU/L)	Measurements outside the CoA
2017-S1 (n = 10)	11.0 (L)	1 (L)
	47.0 (H)	1 (H)
2018-S1 (n = 35)	10.0 (L)	1 (L)
	54.0 (H)	4 (H)
2018-S2 (n = 33)	22.0 (L)	2 (L)
	89.0 (H)	0 (H)
2019-S1 (n = 52)	13.0 (L)	4 (L)
	35.0 (H)	3 (H)
2019-S2 (n = 49)	6.6 (L)	23 (L)
	62.1 (H)	7 (H)
2020-S1a (n = 34)	6.3 (L)	13 (L)
	60.1 (H)	4 (H)
2020-S1b (n = 26)	9.9 (L)	3 (L)
	48.2 (H)	2 (H)
2020-S2a (n = 32)	12.1 (L)	1 (L)
	33.5 (H)	1 (H)
2020-S2b (n = 28)	21.7 (L)	3 (L)
	86.1 (H)	2 (H)
2021-S1 (n = 84)	10.2 (L)	8 (L)
	22.6 (H)	13 (H)
2021-S2 (n = 75)	16.2 (L)	2 (L)
	66.1 (H)	9 (H)
2022-S1 (n = 93)	17.3 (L)	6 (L)
	32.9 (H)	5 (H)
2022-S2 (n = 77)	10.2 (L)	8 (L)
	16.2 (H)	7 (H)

subsequent years, the number of participants increased to an annual average of 85 laboratories in 2022 (Figure 1; Table 1). The overall median (IQR) percentage of participants not meeting the CoA was 9.5% (6.1%–10.7%) for sample L and 9.1% (5.8%–11.8%) for sample H. Relatively high rates (46.9% and 38.2%, respectively) of measurements outside the CoA for sample L were observed for 2019-S2 and 2020-S1a (Figure 1; Table 1). The interlaboratory variation was determined by calculating the MAD/median ratio for each survey. The overall MAD/median ratio (median; IQR) was 11.0% (7.5%–13.1%) (sample L) and 9.9% (8.8%–10.6%) (sample H) but showed an unusual peak for 2019-S2 at 25.0% for sample L, which is in line with the low passing rate for this survey (Figure 1).

The results were also evaluated based on the immunological methods used by the laboratories (Table 2). Scatterings for the individual methods are quite low when considered in relation to the overall distribution of the data. Overall, the method-specific data distributions were mostly within the quartiles of the total data distribution. In some cases, the value distribution for ELISA shifted upwards, especially for the less concentrated samples (sample L) between 2019-S2 and 2020-S2b but also for sample H 2020-S2a and 2021-S1 (Figure 2). With 824 measurements for samples L and H combined, CLIA was the most frequently used method in every survey. LEIA had the lowest frequency, with 118 total observations. ELISA was used 132 times.

Regarding the reagent manufacturer-dependent data analysis, the most frequently used manufacturer was DG, with 846 measurements for sample L and sample H combined (Table 3). Manufacturer BE was used 158 times. IB was used the least (n = 70). IB showed a tendency for higher values, and upward shifts could be observed in some surveys, especially for sample L between 2019-S2 and 2021-S1, but also for sample H in the 2019 and 2021 surveys (Figure 3). In some cases, BE tended towards values in the lower range of the overall distribution and, in some cases, even outside the lower quartile. One shift outside the upper quartile could

TABLE 2 Method-dependent and total median (interquartile range; IQR; IU/L) and respective frequencies in each survey (S) from 2017 to 2022 for sample L (L) and sample H (H).

Survey	Median (IQR; IU/L) Frequency				
	CLIA	ELISA	LEIA	other	total
2017-S1	10.8 (10.1–11.1) (L)	10.5 (-) (L)	12.2 (-) (L)	10.7 (10.0–11.1) (L)	10.7 (10.4–11.2) (L)
	47.9 (45.3–48.4) (H)	41.0 (-) (H)	46.5 (-) (H)	49.1 (47.6–51.9) (H)	47.9 (44.8–48.9) (H)
	n = 4	n = 1	n = 1	n = 4	n = 10
2018-S1	9.7 (9.4–10.1) (L)	10.8 (9.9–11.2) (L)	9.9 (9.8–10.0) (L)	9.6 (9.4–10.3) (L)	9.7 (9.5–10.1) (L)
	53.6 (53.0–58.2) (H)	46.4 (42.3–51.6) (H)	52.6 (51.5–55.3) (H)	51.2 (46.9–60.8) (H)	53.5 (50.8–58.1) (H)
	n = 21	n = 3	n = 3	n = 8	n = 35
2018-S2	21.9 (21.0–22.8) (L)	24.5 (-) (L)	22.5 (21.0–23.5) (L)	22.0 (21.2–25.3) (L)	21.9 (21.0–23.0) (L)
	89.9 (84.2–95.5) (H)	95.7 (-) (H)	83.8 (82.8–86.3) (H)	89.8 (85.1–94.7) (H)	89.6 (83.8–95.5) (H)
	n = 21	n = 1	n = 3	n = 8	n = 33
2019-S1	12.9 (12.2–13.3) (L)	13.4 (12.8–15.2) (L)	13.2 (12.7–13.8) (L)	13.0 (12.1–13.6) (L)	12.9 (12.2–13.6) (L)
	34.6 (32.7–35.9) (H)	36.4 (33.5–37.9) (H)	35.5 (34.0–37.1) (H)	36.9 (35.3–38.0) (H)	35.2 (32.9–36.9) (H)
	n = 32	n = 7	n = 4	n = 9	n = 52
2019-S2	6.1 (5.6–7.1) (L)	9.7 (9.1–10.3) (L)	6.4 (6.1–6.8) (L)	5.3 (4.0–8.2) (L)	6.4 (5.5–7.5) (L)
	60.6 (57.6–64.6) (H)	68.8 (64.9–70.6) (H)	63.2 (63.1–63.4) (H)	63.8 (56.3–68.9) (H)	62.2 (57.6–66.6) (H)
	n = 32	n = 6	n = 2	n = 9	n = 49
2020-S1a	6.1 (5.4–6.4) (L)	9.6 (8.4–12.4) (L)	5.7 (5.7–6.8) (L)	6.7 (5.7–8.3) (L)	6.2 (5.7–6.8) (L)
	61.8 (56.2–64.2) (H)	60.6 (54.5–68.8) (H)	58.2 (55.5–59.7) (H)	60.9 (57.7–63.5) (H)	60.5 (55.9–64.2) (H)
	n = 21	n = 4	n = 5	n = 4	n = 34
2020-S1b	9.2 (8.6–9.7) (L)	10.3 (9.7–11.3) (L)	9.1 (-) (L)	10.6 (9.5–14.4) (L)	9.4 (8.6–10.2) (L)
	50.0 (47.6–51.9) (H)	45.4 (38.3–50.8) (H)	50.2 (-) (H)	46.0 (43.5–51.3) (H)	49.3 (44.6–51.8) (H)
	n = 17	n = 5	n = 1	n = 3	n = 26
2020-S2a	12.0 (11.7–12.5) (L)	13.4 (13.2–13.5) (L)	11.7 (11.3–12.3) (L)	11.7 (11.1–13.0) (L)	12.0 (11.7–12.9) (L)
	33.7 (32.1–34.5) (H)	37.5 (37.1–38.0) (H)	31.8 (31.6–33.2) (H)	31.6 (30.5–39.6) (H)	33.5 (31.8–34.8) (H)
	n = 20	n = 2	n = 5	n = 5	n = 32
2020-S2b	21.1 (20.0–22.2) (L)	24.4 (22.8–26.7) (L)	-	21.8 (20.9–24.4) (L)	21.7 (20.0–23.5) (L)
	87.6 (83.3–91.9) (H)	86.8 (77.7–94.7) (H)	-	83.8 (73.1–88.3) (H)	86.5 (79.9–92.6) (H)
	n = 21	n = 4	n = 0	n = 3	n = 28
2021-S1	10.0 (9.4–10.6) (L)	12.0 (9.4–15.1) (L)	10.1 (9.5–11.3) (L)	10.5 (9.0–11.8) (L)	10.0 (9.4–10.9) (L)
	22.0 (21.0–23.8) (H)	29.5 (28.3–32.0) (H)	23.1 (22.4–23.9) (H)	23.0 (19.4–25.5) (H)	22.6 (21.1–24.2) (H)
	n = 56	n = 9	n = 10	n = 9	n = 84
2021-S2	16.6 (14.9–17.7) (L)	15.4 (14.1–16.5) (L)	16.8 (16.0–17.7) (L)	14.8 (14.0–16.0) (L)	16.1 (14.8–17.6) (L)
	65.0 (60.5–69.3) (H)	75.9 (68.0–80.3) (H)	69.0 (67.9–72.4) (H)	66.9 (62.5–68.2) (H)	67.0 (62.5–70.8) (H)
	n = 50	n = 8	n = 8	n = 9	n = 75
2022-S1	17.4 (16.3–18.3) (L)	17.6 (15.8–20.4) (L)	16.8 (16.2–17.6) (L)	16.9 (15.8–18.2) (L)	17.3 (16.1–18.3) (L)
	33.2 (31.3–35.5) (H)	31.2 (26.6–33.3) (H)	34.0 (32.5–35.4) (H)	33.3 (29.3–34.6) (H)	33.2 (31.0–35.2) (H)
	n = 64	n = 9	n = 10	n = 10	n = 93

(Continued on following page)

TABLE 2 (*Continued*) Method-dependent and total median (interquartile range; IQR; IU/L) and respective frequencies in each survey (S) from 2017 to 2022 for sample L (L) and sample H (H).

Survey	Median (IQR; IU/L) Frequency				
	CLIA	ELISA	LEIA	other	total
2022-S2	9.8 (9.0–10.6) (L)	9.9 (8.6–11.6) (L)	10.4 (9.7–11.4) (L)	9.5 (8.6–10.4) (L)	9.9 (8.9–10.6) (L)
	16.8 (15.3–18.1) (H)	16.1 (15.2–18.0) (H)	18.0 (16.6–18.9) (H)	16.0 (14.6–17.4) (H)	16.8 (15.3–18.1) (H)
	n = 53	n = 7	n = 7	n = 10	n = 77

also be seen for sample L in 2020-S1a. Manufacturer DG mostly showed values in the mid-range of the overall data.

4 Discussion

This study summarizes quantitative EQA results for EPO determination conducted between 2017 and 2022. The MAD/median ratio was below 15% in almost every case. Survey 2019-S2 showed higher values at 25.0% for sample L. Also, some immunological methods and reagent manufacturers showed variability in measurement outcomes to some extent. These findings should also be placed in relation to their clinical relevance. EPO determination is mainly a diagnosis of exclusion to identify, for example, chronic kidney disease as the cause of anemia. Therefore, the focus is on the concentration of EPO in relation to other anemia markers rather than on the exact prevailing EPO concentration. Low EPO concentrations in the blood, in combination with hemoglobin concentrations below 13.0 g/dL (adult males) and 12.0 g/dL (non-menstruating females), may indicate a renal cause (Lankhorst and Wish, 2010). Non-renal anemia usually results in increased EPO levels, and, in severe cases, an increase of up to 1000-fold can be reached (Artunc and Risler, 2007; Higgs et al., 2015). Hence, measurement deviations may be, to a small extent, clinically less critical if the EPO value is considered in relation to the relevant biomarkers. Nevertheless, clinical laboratories should always strive for the highest measurement precision so that patient safety, as the highest priority, is never compromised. To this date, further investigation is needed to get clear statements on quality specifications for EPO measurement variation.

Scattering in the EPO levels of the investigated immunological methods and reagent manufacturers could be observed in some cases. Immunoassays have an analytical error rate of 0.4%–4% (Ismail, 2017). This can be attributed to exogenous factors such as variability in sample pipetting and other handling errors or systematic exogenous error sources such as calibration errors (Sturgeon and Viljoen, 2011). Furthermore, interfering factors, such as the reagents used, have been known to affect measurement outcomes (Alhajj and Farhana, 2022). There also may be excessive non-specific binding of the antibody or antigen in the assay performed (Gan and Patel, 2013). It is known that the imprecision of EPO quantification immunoassays depends on the concentration (Marsden, 2006). Especially for the reagent manufacturer IB, scatter could be observed at median sample concentrations of 10 IU/L or less. This manufacturer was only used in combination with the ELISA and “other” method collective. The concentration range of the calibration curve is 10.7–469 IU/L of the commercially available ELISA kit from this manufacturer, according to the manufacturer’s website (IBL International GmbH, 2023). Thus, the EPO concentration in the samples might have been too close to the detection limit of the assay. However, due to the comparably small number of IB applications, more measurements would be needed to corroborate this assumption. Compared to IB, the lowest limit of detection for the manufacturer DG device Immulite 2000 was found to be 0.16 IU/L, with the manufacturer’s recommended detection limit being 0.24 IU/L (Benson et al., 2000). The lowest limit of detection for the DG device Advia Centaur Systems is given at 0.75 IU/L (Siemens Healthcare Diagnostics Inc, 2019). The dynamic range of the BE family of Access Immunoassay Systems EPO assays could be determined at 0.6–750 IU/L (Retka et al., 2005; Beckman Coulter, Inc., 2023). Marsden et al. compared different EPO ELISA test kits with radioimmunoassay as a reference test. One kit from the manufacturer IB was also included in that comparison and showed a slight positive bias compared to the reference method. Even though Marsden et al. was conducted in 1999, and no radioimmunoassay was used in the present study, these results are in line with some observed upward shifts for this manufacturer (Marsden et al., 1999).

In some cases, slight fluctuations were also observed for BE. Owen and Roberts compared the test performance of the Access 2 device of this manufacturer with the Immulite 2000 device by the manufacturer DG and obtained comparably good results with both manufacturers (Owen and Roberts, 2011). As the sample sizes for both manufacturers were the same in the study mentioned (n = 101) compared to the extremely varying frequencies of use in this EQA, the results obtained here do not yet indicate a clear difference in the measurement range of the two methods. Owen and Roberts also compared the two manufacturers DG and BE in terms of cross-reactivity with recombinant EPO preparations and found that both differed considerably in the measurement results of samples spiked with Epoetin alfa and Darbepoetin alfa, as the values for BE were in a much higher range—109 IU/L higher and 242 IU/L higher than DG, respectively (Owen and Roberts, 2011). Because the samples used for these EQA surveys were sometimes spiked, differences in cross-reactivity with recombinant EPO as the cause of variability cannot be safely excluded.

The manufacturer DG was used most frequently by the EQA participants in this work. A study by Abellan et al. from 2004 compares the Immulite 2000 system from DG, which is based on CLIA, with an ELISA kit by a different manufacturer that was not used by any participant in the present study. The DG device showed better intra-laboratory precision and a lower variation in the interlaboratory comparison. Both immunoassay methods correlated well, although ELISA tended to show lower

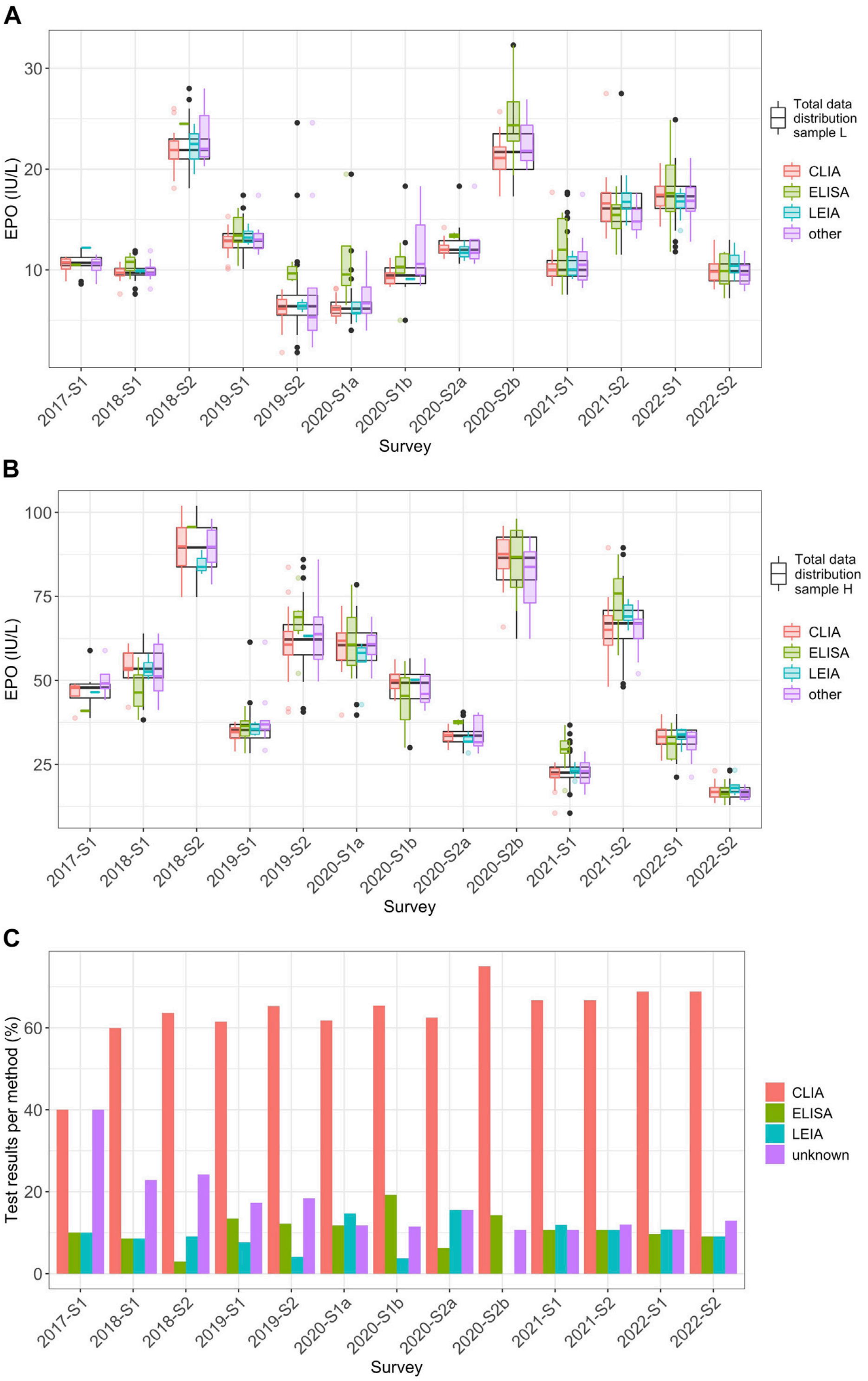


FIGURE 2
Method-dependent analysis of EQA results for EPO levels from 2017 to 2022 **(A)** Distribution of the EPO measurement results (IU/L) for the individual methods CLIA (red), ELISA (green), LEIA (turquoise), and "other" (violet) in relation to the overall distribution of all measured values in the individual surveys (black) for sample L from 2017 to 2022. In this plot, whiskers span 1.5 times the IQR above and below the box, capturing the middle 50% of the data. The red, green, turquoise, violet and black dots mark outliers, which are defined as observations that exceed 1.5 times the IQR from either edge of the box. **(B)** The same consideration used for **(A)** but for sample H. **(C)** Percentage of the frequencies for the respective measurement methods of the total of all measurements per survey per sample.

TABLE 3 Manufacturer-dependent and total median interquartile range (IQR; IU/L) and respective frequencies in each survey (S) from 2017 to 2022 for sample L (L) and sample H (H).

	Median ± IQR (IU/L) Frequency				
Survey	BE	DG	IB	other	total
2017-S1	9.6 (9.2–10.0) (L)	11.3 (10.9–11.5) (L)	8.6 (-) (L)	10.8 (10.6–10.9) (L)	10.7 (10.4–11.2) (L)
	48.8 (43.8–53.9) (H)	48.2 (46.5–49.0) (H)	48.7 (-) (H)	44.2 (42.6–45.9) (H)	47.9 (44.8–48.9) (H)
	n = 2	n = 5	n = 1	n = 2	n = 10
2018-S1	9.8 (8.7–10.8) (L)	9.7 (9.6–10.0) (L)	9.4 (8.8–10.1) (L)	10.2 (9.7–11.1) (L)	9.7 (9.5–10.1) (L)
	53.0 (47.5–58.5) (H)	53.6 (52.2–58.2) (H)	58.4 (57.6–59.1) (H)	41.8 (41.2–46.4) (H)	53.5 (50.8–58.1) (H)
	n = 2	n = 26	n = 2	n = 5	n = 35
2018-S2	20.1 (19.4–22.5) (L)	22.0 (21.2–22.9) (L)	22.4 (21.4–23.4) (L)	22.8 (22.2–24.9) (L)	21.9 (21.0–23.0) (L)
	81.2 (78.9–85.2) (H)	89.9 (86.5–95.6) (H)	87.2 (82.9–91.4) (H)	89.6 (87.0–93.8) (H)	89.6 (83.8–95.5) (H)
	n = 4	n = 24	n = 2	n = 3	n = 33
2019-S1	10.7 (10.2–12.7) (L)	13.0 (12.5–13.6) (L)	13.2 (12.3–14.0) (L)	12.9 (12.4–13.4) (L)	12.9 (12.2–13.6) (L)
	30.9 (29.7–33.0) (H)	35.3 (33.6–36.8) (H)	51.9 (47.1–56.6) (H)	35.1 (34.1–37.1) (H)	35.2 (32.9–36.9) (H)
	n = 4	n = 37	n = 2	n = 9	n = 52
2019-S2	8.0 (7.6–8.1) (L)	5.9 (5.1–6.8) (L)	17.7 (14.3–21.2) (L)	8.9 (7.6–9.2) (L)	6.4 (5.5–7.5) (L)
	58.8 (57.5–60.7) (H)	62.5 (58.5–66.4) (H)	72.6 (68.6–76.5) (H)	63.8 (55.2–65.9) (H)	62.2 (57.6–66.6) (H)
	n = 6	n = 34	n = 2	n = 7	n = 49
2020-S1a	9.8 (8.8–10.9) (L)	6.0 (5.5–6.8) (L)	-	6.5 (6.3–8.4) (L)	6.2 (5.7–6.8) (L)
	63.5 (60.8–66.3) (H)	59.9 (54.5–62.9) (H)	-	62.9 (58.4–65.2) (H)	60.5 (55.9–64.2) (H)
	n = 2	n = 26	-	n = 6	n = 34
2020-S1b	8.6 (8.5–8.7) (L)	9.2 (8.7–9.6) (L)	14.8 (13.1–16.6) (L)	10.4 (10.3–10.6) (L)	9.4 (8.6–10.2) (L)
	44.1 (43.9–44.3) (H)	50.1 (48.0–51.8) (H)	56.2 (55.9–56.4) (H)	45.4 (41.0–50.8) (H)	49.3 (44.6–51.8) (H)
	n = 5	n = 14	n = 2	n = 5	n = 26
2020-S2a	11.8 (11.5–13.6) (L)	12.0 (11.7–12.5) (L)	-	13.1 (13.0–13.1) (L)	12.0 (11.7–12.9) (L)
	34.6 (34.3–36.2) (H)	33.1 (31.6–34.2) (H)	-	38.2 (37.4–38.9) (H)	33.5 (31.8–34.8) (H)
	n = 4	n = 26	n = 0	n = 2	n = 32
2020-S2b	20.1 (19.7–21.0) (L)	21.8 (20.6–22.2) (L)	25.4 (24.6–26.1) (L)	21.4 (19.2–25.7) (L)	21.7 (20.0–23.5) (L)
	83.3 (79.7–84.4) (H)	91.8 (86.5–93.0) (H)	71.2 (66.8–75.6) (H)	82.2 (74.8–90.6) (H)	86.5 (79.9–92.6) (H)
	n = 7	n = 15	n = 2	n = 4	n = 28
2021-S1	9.0 (8.8–9.1) (L)	10.2 (9.5–10.8) (L)	14.8 (13.1–16.1) (L)	9.5 (9.2–10.5) (L)	10.0 (9.4–10.9) (L)
	21.6 (21.2–22.1) (H)	22.5 (21.0–24.0) (H)	31.5 (28.9–34.8) (H)	23.3 (21.6–27.7) (H)	22.6 (21.1–24.2) (H)
	n = 8	n = 58	n = 4	n = 14	n = 84
2021-S2	14.8 (14.4–15.0) (L)	17.1 (15.9–18.1) (L)	16.4 (14.2–16.7) (L)	13.8 (13.2–15.4) (L)	16.1 (14.8–17.6) (L)
	62.9 (58.6–64.4) (H)	67.5 (64.3–70.8) (H)	77.5 (73.7–80.0) (H)	57.5 (52.7–64.9) (H)	67.0 (62.5–70.8) (H)
	n = 12	n = 47	n = 5	n = 11	n = 75
2022-S1	17.8 (16.4–18.4) (L)	17.3 (16.1–18.2) (L)	16.5 (12.6–18.2) (L)	17.7 (16.2–19.1) (L)	17.3 (16.1–18.3) (L)
	30.3 (28.4–31.0) (H)	34.0 (32.5–35.5) (H)	27.1 (26.4–33.0) (H)	31.4 (28.5–35.1) (H)	33.2 (31.0–35.2) (H)
	n = 11	n = 63	n = 7	n = 12	n = 93

(Continued on following page)

TABLE 3 (*Continued*) Manufacturer-dependent and total median interquartile range (IQR; IU/L) and respective frequencies in each survey (S) from 2017 to 2022 for sample L (L) and sample H (H).

	Median ± IQR (IU/L) Frequency				
Survey	BE	DG	IB	other	total
2022-S2	9.0 (8.6–9.6) (L)	10.2 (9.7–10.8) (L)	9.2 (8.4–11.3) (L)	9.1 (8.5–9.7) (L)	9.9 (8.9–10.6) (L)
	14.6 (14.3–16.0) (H)	17.5 (16.5–18.2) (H)	15.4 (14.5–17.1) (H)	15.9 (14.7–17.4) (H)	16.8 (15.3–18.1) (H)
	n = 12	n = 48	n = 6	n = 11	n = 77

values (Abellan et al., 2004). In the methodological comparison of the present study, some cases were observed in which ELISA tended to show higher values than CLIA and LEIA, which contrasts with the tendency observed in the mentioned article.

Because there is not yet any reference method for quantitative EPO determination, no valid statement can be made as to which method or which manufacturer offers the highest precision. External quality controls are, therefore, even more important when comparing the measuring ranges of the laboratories and the methodology. Methodological comparisons require representative sample sizes, which are partially not yet given due to the low frequency of use in some cases. Because the number of participants in the EPO EQA has been increasing, more specific comparisons might be made in future studies.

It should also be noted that the standards used for the IBL-ELISA were calibrated against the first international erythropoietin standard (87/684) (IBL International GmbH, 2022 National Institute for Biological Standards and Control, 2008). The calibrator of the Immulite 2000 by manufacturer BE and the devices used in this study from manufacturer DG are traceable to the second international erythropoietin standard (67/343) (Owen and Roberts, 2011; Beckman Coulter, Inc., 2020). The second international standard is derived from urine but is used to calibrate detection in human serum or plasma (National Institute for Biological Standards and Control, 2013). It remains questionable whether accurate results can be obtained in blood if the calibrators of the assays are traceable to a standard from a completely different matrix. The Siemens Advia Centaur device from manufacturer BE, which was used by some participants in this study, is traceable to the second international standard and the third international erythropoietin standard (11/170), which is mainly based on a recombinant EPO preparation (National Institute for Biological Standards and Control, 2012; Siemens Healthcare Diagnostics Inc, 2019).

EQAs may not be passed for different reasons, most of which can be attributed to human error, such as sample mix-ups or errors during the reconstitution process. Li et al. found that potential reasons for not passing EQAs can, for example, be due to errors in the management of the measurement results, such as transcription errors or reporting of incorrect units, which were also noticed in this work. However, technical errors, such as calibration problems, were described as the main reason (Li et al., 2019). To successfully complete the EQA, it is important that participants follow the details of the test scheme and apply good laboratory practices, like checking the methods for quality and ensuring that the staff is adequately trained (Edson et al., 2007). Two surveys (2019-S2 and 2020-S1a) did stand out with a particularly high failure rate and high interlaboratory variation for sample L. The same batch of sample sets was used in these two surveys. This suggests that there might be interfering components in this batch for sample L. This may be due to unusually high concentrations of regular serum components prevailing in the sample, leading to falsely high or falsely low results (Sequeira, 2019). Insufficient commutability of the sample may also have negatively impacted the test performance. It is often not possible to use authentic clinical samples in the context of proficiency testing. However, artificially generated samples do not always mirror the patient samples that are routinely examined in laboratories (Laudus et al., 2022). In the EQA surveys performed, samples were sent to participants in lyophilized form. The samples used in 2019-S2 and 2020-S1a were not spiked with recombinant EPO, but other samples used in this study were. Both sample preparation and sublimation have been described as possible influencing factors (Vesper et al., 2007; Miller et al., 2011). As mentioned above, there can also be differences in cross-reactivity with recombinant EPO preparations depending on the assay manufacturer (Owen and Roberts, 2011).

The study had the following limitations: The exact isoform of recombinant EPO spiked into some of the samples is unknown. This makes it difficult to draw conclusions about any possible cross-reactivity in the samples. It should also be reiterated here that commutability studies of the EQA samples have not yet been carried out, so a possible influence of the sample preparation on EPO detection is not known. Whether the test performance is affected by the sample itself should be evaluated. As mentioned above, there is no validated reference method for quantitative EPO detection. Accordingly, no analytical target value can be determined for the evaluation of the EQA, and the robust mean value must be used as the target value for evaluation, which is a common practice. The most represented method or manufacturer also has the strongest influence on the overall mean. Because the true value is unknown, this can lead to biases in the evaluation to an unknown extent (Kristensen and Meijer, 2017). Furthermore, it is not possible to include the exact specifications given by the manufacturer for each method at any given time, as the corresponding reagent kits and batches are not known. The EQA is also intended to provide an overall picture of the analyses rather than comparing individual kits and batches.

However, the results presented in this study are of importance despite the limitations mentioned, as this is the first longitudinal evaluation of EPO EQA data to date. Medical laboratories should always aim to keep their measurement quality at the highest standard, and this work can be used to reflect on the institution's

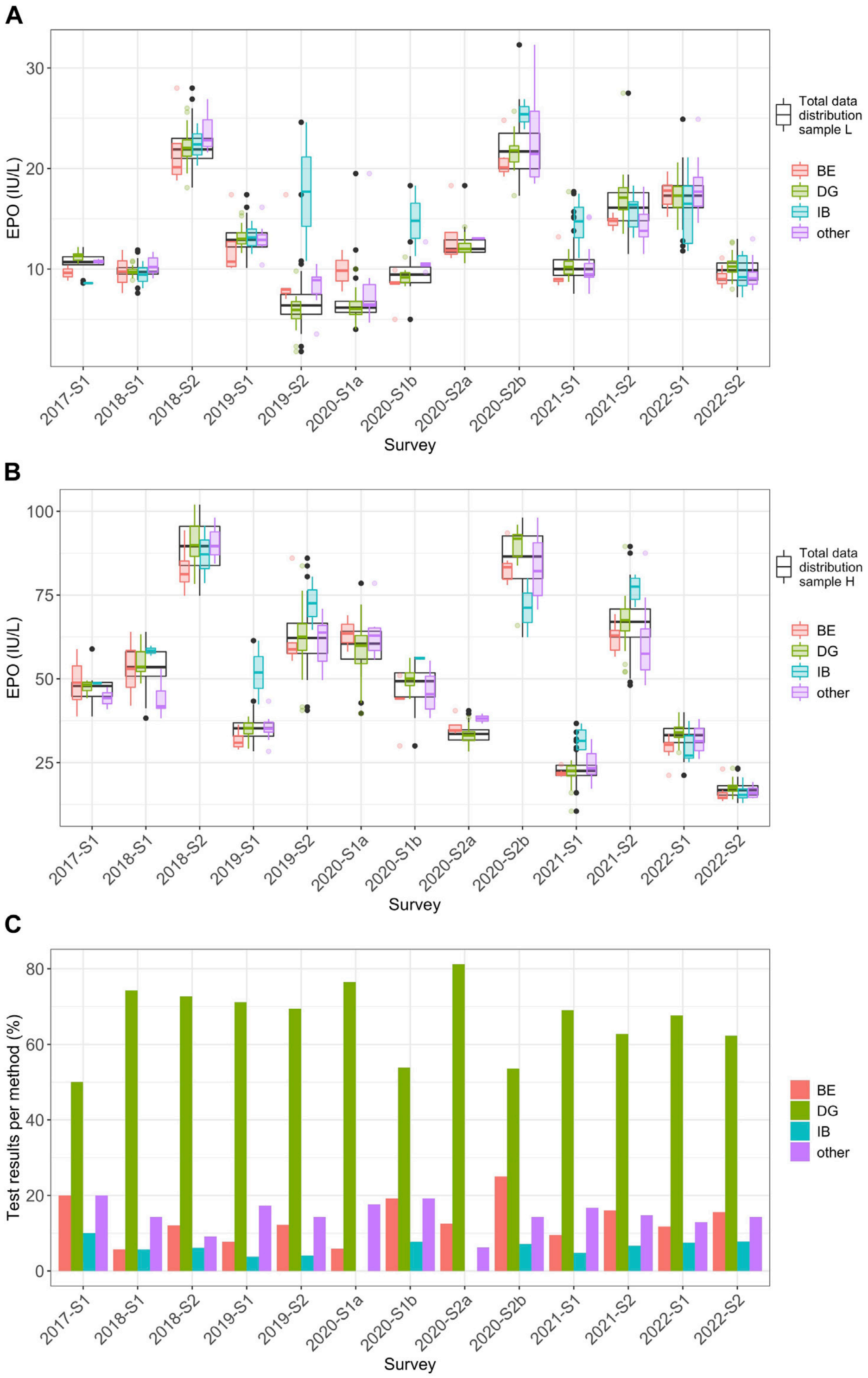


FIGURE 3
Manufacturer-dependent analysis of EQA results for EPO levels from 2017 to 2022. **(A)** Distribution of the EPO measurement results (IU/L) for the individual reagent manufacturers BE (red), DG (green), IB (turquoise), and "other" (violet) in relation to the overall distribution of all measured values in the individual surveys (black) for sample L from 2017 to 2022. In this plot, whiskers span 1.5 times the IQR above and below the box, capturing the middle 50% of the data. The red, green, turquoise, violet and black dots mark outliers, which are defined as observations that exceed 1.5 times the IQR from either edge of the box. **(B)** The same consideration used in **(A)** but for sample H. **(C)** Percentage of the frequencies for the respective manufacturers of the total of all measurements per survey per sample.

methodology and to see how their detection method or the assay manufacturer used performs in relation to others.

5 Conclusion

This work shows that variations in laboratory results and in methodological terms for quantitative EPO determination do persist to some degree, and knowledge about sources of errors is vital in order to optimize measurement quality and thus ensure patient safety. However, in terms of clinical relevance, small deviations might be considered less critical for the diagnostic assessment and the resulting therapeutic consequence in patients because, in anemia diagnostics, the level of EPO in combination with other relevant biomarkers is of decisive importance. Thresholds for maximum acceptable variation in EPO measurement quality and their clinical consequences should be further investigated in the future.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

LT: Formal Analysis, Visualization, Writing–original draft. NW: Writing–review and editing. LV: Writing–review and editing. IS: Supervision, Writing–review and editing. MT: Writing–review and editing. FW: Supervision, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2024.1390079/full#supplementary-material

SUPPLEMENTARY TABLE 1:
Overview of the collectives considered in the data analysis and corresponding information on the reagent manufacturers and devices used.

SUPPLEMENTARY TABLE 2:
Raw data.

References

Abellan, R., Ventura, R., Pichini, S., Remacha, A. F., Pascual, J. A., Pacifici, R., et al. (2004). Evaluation of immunoassays for the measurement of erythropoietin (EPO) as an indirect biomarker of recombinant human EPO misuse in sport. *J. Pharm. Biomed. Analysis* 35, 1169–1177. doi:10.1016/j.jpba.2004.02.001

Alhajj, M., and Farhana, A. (2022). "Enzyme linked immunosorbent assay," in *StatPearls* (Treasure Island (FL): StatPearls Publishing). Available at: https://www.ncbi.nlm.nih.gov/books/NBK555922/(Accessed March 8, 2022).

Artunc, F., and Risler, T. (2007). Serum erythropoietin concentrations and responses to anaemia in patients with or without chronic kidney disease. *Nephrol. Dial. Transplant.* 22, 2900–2908. doi:10.1093/ndt/gfm316

Beckman Coulter, Inc (2020). "ACCESS immunoassay systems, instructions for use," in *Access EPO calibrators erythropoietin, REF A16365*. Available at: https://www.beckmancoulter.com/download/file/phxB50133J-EN_US/B50133J?type=pdf (Accessed January 30, 2024).

Beckman Coulter, Inc (2023). Beckman coulter ACCESS immunoassay systems. *Access EPO Erythropoietin - Instr. Use*. Available at: https://www.beckmancoulter.com/download/file/phxC94197A-EN_US/C94197A?type=pdf (Accessed January 30, 2024).

Benson, E. W., Hardy, R., Chaffin, C., Robinson, C. A., and Konrad, R. J. (2000). New automated chemiluminescent assay for erythropoietin. *J. Clin. Laboratory Analysis* 14, 271–273. doi:10.1002/1098-2825(20001212)14:6<271::AID-JCLA4>3.0.CO;2-8

Bundesärztekammer (2022) *Richtlinie der Bundesärztekammer zur Qualitätssicherung laboratoriumsmedizinischer Untersuchungen Deutsches Ärzteblatt Online*. doi:10.3238/arztebl.2019.rili_baek_QS_Labor20222511

Bunn, H. F. (2013). Erythropoietin. *Cold Spring Harb. Perspect. Med.* 3, a011619. doi:10.1101/cshperspect.a011619

De La Salle, B., Meijer, P., Thomas, A., and Simundic, A.-M. (2017). Special issue on external quality assessment in laboratory medicine – current challenges and future trends. *Biochem. Med.* 27, 19–22. doi:10.11613/BM.2017.003

DIN EN ISO 15189:2023-03 (2023) *Medizinische Laboratorien - Anforderungen an die Qualität und Kompetenz (ISO 15189:2022); Deutsche Fassung EN ISO 15189:2022.*

Edson, D. C., Russell, D., and Massey, L. D. (2007). Proficiency testing: a guide to maintaining successful performance. *Lab. Med.* 38, 184–186. doi:10.1309/B0GAEHHQ765RD3V7

Favaloro, E. J., Jennings, I., Olson, J., Van Cott, E. M., Bonar, R., Gosselin, R., et al. (2018). Towards harmonization of external quality assessment/proficiency testing in hemostasis. *Clin. Chem. Laboratory Med. (CCLM)* 57, 115–126. doi:10.1515/cclm-2018-0077

Fried, W. (2009). Erythropoietin and erythropoiesis. *Exp. Hematol.* 37, 1007–1015. doi:10.1016/j.exphem.2009.05.010

Gan, S. D., and Patel, K. R. (2013). Enzyme immunoassay and enzyme-linked immunosorbent assay. *J. Investigative Dermatology* 133, e12–e13. doi:10.1038/jid.2013.287

Higgs, D. R., Roy, N., and Hay, D. (2015). "Erythropoiesis," in *Postgraduate haematology* Editors A. V. Hoffbrand, D. R. Higgs, D. M. Keeling, and A. B. Mehta (New Jersey, United States: Wiley), 11–20. doi:10.1002/9781118853771.ch2

IBL International GmbH (2022). Erythropoietin ELISA, Enzyme immunoassay for the quantitative determination of Erythropoietin (EPO) in human serum. Available at: https://www.ibl-international.com/media/mageworx/downloads/attachment/file/n/m/nm56011_ifu_us_en_epo_elisa_2022-05_sym9.pdf.

IBL International GmbH (2023). Erythropoietin (EPO) ELISA - quality control certificate. Available at: https://ibl-international.com/media/mageworx/downloads/attachment/file/n/m/nm56011_qc_3530__240229.pdf (Accessed September 26, 2023).

Ismail, A. A. (2017). When laboratory tests can mislead even when they appear plausible. *Clin. Med.* 17, 329–332. doi:10.7861/clinmedicine.17-4-329

ISO13528:2015 (2020) *Statistical methods for use in proficiency testing by interlaboratory comparison. (ISO 13528:2015, Corrected version 2016-10-15), I.O.f. Standardization, Editor. Geneva, Switzerland. Section 3*.

Jelkmann, W. (2011). Regulation of erythropoietin production: erythropoietin production. *J. Physiology* 589, 1251–1258. doi:10.1113/jphysiol.2010.195057

Kristensen, G. B. B., and Meijer, P. (2017). Interpretation of EQA results and EQA-based trouble shooting. *Biochem. Med.* 27, 49–62. doi:10.11613/BM.2017.007

Lankhorst, C. E., and Wish, J. B. (2010). Anemia in renal disease: diagnosis and management. *Blood Rev.* 24, 39–47. doi:10.1016/j.blre.2009.09.001

Laudus, N., Nijs, L., Nauwelaers, I., and Dequeker, E. M. C. (2022). The significance of external quality assessment schemes for molecular testing in clinical laboratories. *Cancers* 14, 3686. doi:10.3390/cancers14153686

Li, T., Zhao, H., Zhang, C., Wang, W., He, F., Zhong, K., et al. (2019). Reasons for proficiency testing failures in routine chemistry analysis in China. *Lab. Med.* 50, 103–110. doi:10.1093/labmed/lmy032

Marsden, J. T. (2006). Erythropoietin - measurement and clinical applications. *Ann. Clin. Biochem.* 43, 97–104. doi:10.1258/000456306776021553

Marsden, J. T., Day, P., Ellis, R., Marwah, S., Savage, G., and Sinclair, C. (2006). A sample distribution programme for erythropoietin. *Clin. Laboratory Haematol.* 28, 228–232. doi:10.1111/j.1365-2257.2006.00786.x

Marsden, J. T., Sherwood, R. A., and Peters, T. J. (1999). Evaluation of six erythropoietin kits. *Ann. Clin. Biochem.* 36, 380–387. doi:10.1177/000456329903600312

Michiels, J. J., De Raeve, H., Hebeda, K., Lam, K. H., Berneman, Z., Schroyens, W., et al. (2007). WHO bone marrow features and European clinical, molecular, and pathological (ECMP) criteria for the diagnosis of myeloproliferative disorders. *Leukemia Res.* 31, 1031–1038. doi:10.1016/j.leukres.2007.01.021

Miller, W. G., Jones, G. R., Horowitz, G. L., and Weykamp, C. (2011). Proficiency testing/external quality assessment: current challenges and future directions. *Clin. Chem.* 57, 1670–1680. doi:10.1373/clinchem.2011.168641

National Institute for Biological Standards and Control (2008) *WHO international standard erythropoietin, human recombinant, NIBSC code: 88/574 instructions for use (version 4.0, dated 22/01/2008)*. Available at: https://nibsc.org/documents/ifu/88-574.pdf.

National Institute for Biological Standards and Control (2012) *WHO international standard 3rd WHO international standard for erythropoietin, recombinant, for bioassay NIBSC code: 11/170 instructions for use (version 1.0, dated 30/10/2012)*.

National Institute for Biological Standards and Control (2013) *WHO reference reagent, erythropoietin, human, urinary. 2nd international reference preparation NIBSC code: 67/343, instructions for use, (version 5.0, dated 28/03/2013)*. Available at: https://nibsc.org/documents/ifu/67-343.pdf.

Owen, W. E., and Roberts, W. L. (2011). Performance characteristics of a new Immulite® 2000 system erythropoietin assay. *Clin. Chim. Acta* 412, 480–482. doi:10.1016/j.cca.2010.11.023

Portolés, J., Martín, L., Broseta, J. J., and Cases, A. (2021). Anemia in chronic kidney disease: from pathophysiology and current treatments, to future agents. *Front. Med.* 8, 642296. doi:10.3389/fmed.2021.642296

Retka, S., Haning, J., Fuerstenberg, R., Drai, L., and Mylvaganam, R. (2005). Automated erythropoietin testing on beckman coulter's family of Access® immunoassay systems. *Blood* 106, 3761. doi:10.1182/blood.V106.11.3761.3761

Sciacovelli, L., Secchiero, S., Padoan, A., and Plebani, M. (2018). External quality assessment programs in the context of ISO 15189 accreditation. *Clin. Chem. Laboratory Med. (CCLM)* 56, 1644–1654. doi:10.1515/cclm-2017-1179

Sequeira, S. (2019). An overview on interference in clinical immunoassays: a cause for concern. *Hamdan Med. J.* 12, 158. doi:10.4103/HMJ.HMJ_3_19

Siemens Healthcare Diagnostics Inc (2019) *Erythropoietin - atellica IM analyzer and ADVIA Centaur systems*. Available at: https://marketing.webassets.siemens-healthineers.com/9336e80b0cc00c25/330604c31e8c/30-19-14106-01-76_EPO-Assay_SpecSheet_OUS_FINAL_SNG.pdf.

Sturgeon, C. M., and Viljoen, A. (2011). Analytical error and interference in immunoassay: minimizing risk. *Ann. Clin. Biochem.* 48, 418–432. doi:10.1258/acb.2011.011073

Vesper, H. W., Miller, W. G., and Myers, G. L. (2007). Reference materials and commutability. *Clin. Biochem. Rev.* 28 (4), 139–147.

Wickham, H. (2016) *ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag. Available at: https://ggplot2.tidyverse.org.

Zima, T. (2017). Accreditation of medical laboratories – system, process, benefits for labs. *J. Med. Biochem.* 36, 231–237. doi:10.1515/jomb-2017-0025

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 08 August 2024
DOI 10.3389/fmolb.2024.1401405

Longitudinal evaluation of manufacturer-specific differences for high-sensitive CRP EQA results

Nathalie Weiss[1]*, Laura Vierbaum[1], Marcel Kremser[1], Anne Kaufmann-Stoeck[1], Silke Kappler[1], Silvia Ballert[2], Kathrin Kabrodt[2], Klaus-Peter Hunfeld[1,3] and Ingo Schellenberg[1,2]

[1]INSTAND e.V., Society for Promoting Quality Assurance in Medical Laboratories e.V., Duesseldorf, Germany, [2]Institute of Bioanalytical Sciences (IBAS), Center of Life Sciences, Anhalt University of Applied Sciences, Bernburg, Germany, [3]Medical Faculty, Northwest Medical Centre, Academic Teaching Hospital, Institute for Laboratory Medicine, Microbiology and Infection Control, Goethe University Frankfurt, Frankfurt, Germany

OPEN ACCESS

EDITED BY
Artur Slomka,
Nicolaus Copernicus University in Toruń, Poland

REVIEWED BY
Emina Colak,
University of Belgrade, Serbia
A. Thijs,
Amsterdam University Medical Center, Netherlands
Søren Nielsen,
NordiQC, Denmark

*CORRESPONDENCE
Nathalie Weiss,
✉ weiss@instand-ev.de

RECEIVED 15 March 2024
ACCEPTED 16 July 2024
PUBLISHED 08 August 2024

CITATION
Weiss N, Vierbaum L, Kremser M, Kaufmann-Stoeck A, Kappler S, Ballert S, Kabrodt K, Hunfeld K-P and Schellenberg I (2024), Longitudinal evaluation of manufacturer-specific differences for high-sensitive CRP EQA results. *Front. Mol. Biosci.* 11:1401405. doi: 10.3389/fmolb.2024.1401405

Background: C-reactive protein (CRP) is an established serum biomarker for different pathologies such as tissue injury and inflammatory events. One rising area of interest is the incorporation of low concentrations of CRP, so called high-sensitive (hs-) CRP, in the risk assessment and treatment monitoring of cardiovascular diseases (CVDs). Many research projects and the resulting meta-analyses have reported controversial results for the use of hs-CRP, especially in the risk assessment of CVDs. However, since these analyses used different assays to detect hs-CRP, it is important to assess the current level of assay harmonization.

Methods: This paper analyzes data from 17 external quality assessment (EQA) surveys for hs-CRP conducted worldwide between 2018 and 2023. Each EQA survey consisted of two blinded samples. In 2020 the sample material changed from pooled serum to single-donor samples. The aim was to assess the current status of assay harmonization by a manufacturer-based approach, taking into consideration the clinical decision limits for hs-CRP risk-stratification of CVDs as well as the scatter of results.

Results: Our analyses show that harmonization has increased in recent years from median differences of up to 50% to below 20%, with one exception that showed an increasing bias throughout the observed period. After changing sample materials from pools to single-donor samples, the coefficient of variation decreased to below 10% with one exception. Nevertheless, even these differences in the clinical setting could lead to disparate classification of patients depending on the assay used.

Conclusion: While there was a positive trend towards harmonization, meta-analysis of different risk-score publications should stratify their analysis by assay to account for the manufacturer-specific differences observed in this paper.

Furthermore, assays are currently traceable to different international standard preparations, which might have a negative impact on future harmonization.

KEYWORDS

hsCRP, external quality assessment scheme (EQA), proficiency testing (PT), harmonization, cardiovascular diseases

1 Introduction

C-reactive protein (CRP) is an acute-phase protein that is predominantly produced in hepatocytes in response to tissue injury, inflammatory events, acute infection and advanced age (Póvoa et al., 1998; Póvoa, 2002; Pepys and Hirschfield, 2003; Lobo, 2012). After more and more evidence emerged, that cardiovascular diseases (CVDs) such as ischemic stroke and acute myocardial infarction are related to inflammation (Libby et al., 2002), moderately elevated levels of CRP, so called high-sensitive CRP (hs-CRP), gained interest as a potential new biomarker for these diseases. This need was further highlighted by the fact that CVDs account for 17.9 million deaths annually (WHO, 2023). In Germany, they show a rising prevalence (Heidemann et al., 2021) and cost the German healthcare system €56.4 billion, around 13.1% of all German healthcare costs (Statistisches Bundesamt, 2022). Internationally, CVDs are estimated to have cost €282 billion in 2021in the European Union (European Society of Cardiology, 2023), and $219 billion (~€185 billion) in the United States (CDC, 2021). A systematic review of 49 cost-effectiveness studies concluded that early CVD detection and treatment was predominately cost-effective from a healthcare perspective, but it was also noted a lack of standardization in the included studies (Oude Wolcherink et al., 2023). Nevertheless, high hopes were placed on new markers, that allow an early detection of CVD-risk.

While hs-CRP is commonly used in clinical practice as an inflammatory marker in CVD risk assessments (Musunuru et al., 2008; Romero-Cabrera et al., 2022), its actual clinical significance remains controversial. Several meta-analyses show a positive effect of using hs-CRP to detect CVDs (Li et al., 2017; Romero-Cabrera et al., 2022). Further studies support a beneficial outcome for incidences of CVD events when hs-CRP is included in treatment decisions for CVDs (Ridker et al., 2008; Ridker et al., 2017). However, other authors found no or only marginal evidence for an improvement of hs-CRP-supported CVD risk stratification using scoring systems (Shah et al., 2009) and a low predictive utility of hs-CRP (Ahmad et al., 2024). A systematic review by the U.S. Preventive Services Task Force concluded that, based on the studies it reviewed, incorporating hs-CRP in risk stratification would lead to more misclassified individuals and thus overtreatment. It concluded that there is a lack of significantly conclusive clinical trials that evaluate the incremental effect of hs-CRP and other cardiovascular markers for the initiation of preventive therapy (Lin et al., 2018). Due to the inconclusive results, several international clinical guidelines are currently advising against incorporating hs-CRP in the corresponding risk-assessment algorithms (Piepoli et al., 2016; Deutsche Gesellschaft für Allgemeinmedizin und Familienmedizin e.V., 2017; Visseren et al., 2022; Australian Chronic Disease Prevention Alliance, 2023).

One interesting factor that arises is that meta-analyses, like the one performed by Li et al., aggregate data from various publications in which hs-CRP results were obtained using different assays. It is therefore important that the analytical performance of these various diagnostic tests be reliable and, ideally, harmonized regardless of the measurement procedure used. This would allow a more efficient comparison of the results of different CVD studies with respect to the diagnostic properties of hs-CRP. Even though a certified reference material for CRP exists (Charoud-Got et al., 2009), several studies have reported manufacturer-dependent differences for (hs-) CRP detection in serum (Roberts et al., 2000; Roberts et al., 2001; Thanabalasingham et al., 2011; Wojtalewicz et al., 2019; Stevenson et al., 2023).

This study examines the status of current assay-harmonization for hs-CRP based on the longitudinal manufacturer-dependent differences observed in EQA surveys conducted by INSTAND e.V. – Society for Promoting Quality Assurance in Medical Laboratories e.V. between 2018 and 2023.

2 Materials and methods

2.1 Sample materials—properties and preparation

From January 2018 until July 2020, commercially pooled serum samples of 1 mL were used. Starting in September 2020, the material was changed to 0.3 mL samples, mostly from individual blood donors. No stabilizing additives were added (Müller et al., 2009). This study was conducted in accordance with the Statement of the Central Ethics Commission of Germany on the use of human body materials for medical research purposes (no. 20/02/2003; https://www.zentrale-ethikkommission.de). The donor's informed written consent is available for the single donor samples of the project. A positive vote from the ethics committee of Goethe University Frankfurt (Main) has been obtained for samples from voluntary blood donors.

All samples (pools and individual donors) tested negative for HIV, HBV, and HCV. Homogeneity of each sample batch was tested in line with DIN EN ISO/IEC 17043:2023 before the samples were used in the corresponding EQA (International Organization for Standardization [ISO], 2023).

2.2 EQA procedure

The INSTAND EQA scheme for the detection of hs-CRP in serum is offered globally six times a year. It was established due to the rising demand of the marker in routine diagnostics. The

EQA scheme is only mandatory in Germany, if hs-CRP is part of a laboratory's accreditation. For each survey, two blinded samples with different concentrations are sent to the participating laboratories. One sample has hs-CRP levels of around 1 mg/L (low risk) and one of around 2 mg/L (medium risk). Participants are asked to analyze the samples like normal patient samples and to report their quantitative results for hs-CRP to INSTAND's web-based RV-Online platform (http://rv-online.instandev.de) together with information on the respective device, reagent, and method used.

As no reference measurement procedure is currently available, the consensus value of manufacturer-specific collectives, calculated using algorithm A {[International Organization for Standardization (ISO), 2023] Section C3}, serves as the target value for the evaluation of participant results and for laboratory certification. The criterion for passing the EQA was ±30% around the consensus value.

2.3 Data analysis and statistics

Passing rates and participant numbers were evaluated for all EQA surveys conducted between 2018 and 2023 (Supplementary Table S1). Due to the large number of EQA surveys, only the data obtained from the three annual EQAs with the largest number of participants (January, May, and October of each year) were evaluated. This resulted in 17 EQA surveys (Supplementary Table S2). The participating laboratories reported a total of 3,668 results. Results from individual participants that involved sample swaps or reporting errors were excluded from the analysis. This applied to a total of 11 datasets.

The EQA data were analyzed in a manufacturer-dependent manner. Eight manufacturer collectives (number of participants ≥8 in at least half of the surveys) were included in the analysis: Abbott (AB), Beckman Coulter (BE), Beckman Coulter-Olympus (OL), Siemens Healthineers (SI), Siemens-Dade Behring (BW), Siemens-Bayer Health (BG), Siemens-DPC Biermann (DG), and Roche Diagnostics (RO). The distributions of results are shown as box plot diagrams over time. For all boxes, the box covers the 25th percentile, the median and the 75th percentile while the whiskers stretch from the 1st quartile -1.5 ∗ (interquartile range) to the 3rd quartile +1.5 ∗ (interquartile range). The BE collective comprised two manufacturer sub-collectives (BE, OL). Therefore, they were highlighted with the same color but different filling color, since we observed multimodality in several EQA surveys. The same applies to the four manufacturer sub-collectives consolidated under SI (SI, BW, BG, and DG). Detailed information about the (sub)-collectives can be found in Supplementary Table S2. As the clinical decision limits in the literature differed greatly, the decision limits from the Clinical Laboratory Diagnostics Series by Thomas (2023), which are identical with the limits proposed by the U.S. Preventive Services Task Force (2009), were used for this evaluation. These limits are defined as low risk for CVD (below 1 mg/L), medium risk (1–3 mg/L), and high risk (3 mg/L).

The coefficients of variation (CVs) were calculated to quantify the scatter within the manufacturer collectives. Manufacturer-dependent values that scattered further than 1.5 times the inter-quartile range, the width between the 25th and 75th percentiles, were defined as outliers and excluded before the CVs were calculated. These data points are marked in orange in the raw data (Supplementary Table S2).

Harmonization of the different collectives was assessed though a longitudinal comparison of differences in median values.

Basic statistical analyses were performed using jmp 17.2.0 from SAS Institute (Cary, NC, United States). The overlay images were generated using version 2.10.8 of the Gnu image manipulation software.

3 Results

During the observed period, the number of annual participants per survey remained constant with more than 100 laboratories participating in the surveys in January, May, and October and between 56 and 81 laboratories in March, July, and September (Figure 1A). Depending on the survey, between 71% and 89% of participating laboratories were from Germany, between 4% and 23% from other EU countries and 4% to 13% from non-EU countries (Supplementary Table S1).

The passing rate for each EQA survey fluctuated between 78% and 88% from 2018 to May 2020 and rose to over 90% starting in September 2020 (Figure 1B).

The distribution of the EQA results for hs-CRP showed manufacturer-dependent differences particularly for the DG collective, which tended to show notably higher results than the other collectives especially, but not exclusively, in the higher concentration samples (Figure 2B). In EQA samples with hs-CRP levels around the known clinical decision limits of 1 mg/L and 2 mg/L, respectively, single manufacturer collectives stayed below and/or above this decision limit. For example, for the low concentration samples used in October 2023, DG and OL mostly detected values above 1 mg/L, while SI, BW, and BG detected values clearly below 1 mg/L (Figure 2A). In other cases, the scatter of results of single collectives was large enough to span a clinical decision limit, e.g., for AB in the low concentration samples in 2018 and in January and May 2020, and for DG in the high concentration samples sent out in October 2020 as well as in May and October 2021 (Figure 2).

When relative median values are compared, the DG collective had the highest median values of all observed collectives (Figure 3). Interestingly, the differences seem to increase over the observed period, especially for the higher concentration sample. Here the relative difference was over 30% in comparison to all other collectives of the SI group (Figure 3B). While BW and SI also tended to have slightly higher values, this changed in October 2020 when these groups showed lower results than the other manufacturers. At the same time, BG displayed a negative bias down to −35% for the low concentration sample, which was then reduced to the same bias as SI and BW. In general, the relative median values were up to 50% in 2018 and started to be much better aligned in October 2020, essentially only 20% apart, except for the DG collective.

A closer look at the scatter of results shows that many manufacturer collectives had CVs of around 50% and higher for occasional samples [e.g., BG, AB, and RO in January 2018 for the high concentration sample (Figure 4B)]. Beginning in October 2020, the CV of most collectives stayed below 10%, except for DG, which exhibited CVs of over 30% in May and October 2022 for both samples and around 50% in May 2023 for the high concentration sample (Figure 4).

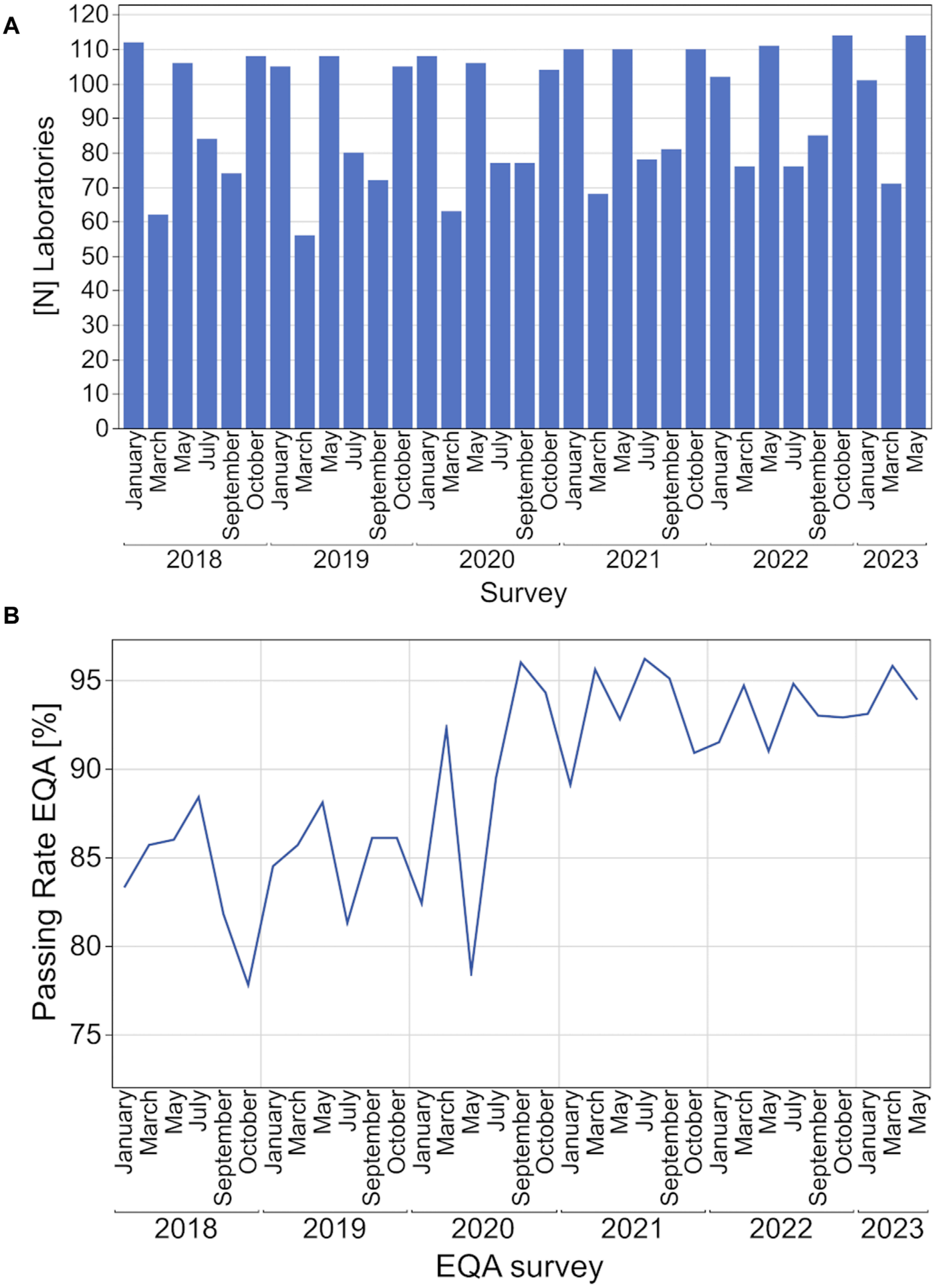


FIGURE 1
Development of participating laboratories **(A, B)** EQA passing rates for hsCRP from 2018 to 2023.

4 Discussion

The significance of the serum-marker CRP has increased in recent decades. While concentrations >5 mg/L are a marker for tissue injury, inflammatory events and acute infection (Póvoa et al., 1998; Póvoa, 2002; Pepys and Hirschfield, 2003; Lobo, 2012), continuous moderately elevated concentrations around and above 2–3 mg/L, so called hs-CRP, have been identified as a possible risk factor for CVDs (U.S. Preventive Services Task Force, 2009; Zhang et al., 2021; Lee and Lee, 2023). This paper assessed the current quality of hs-CRP detection based on EQA data from 2018 to 2023.

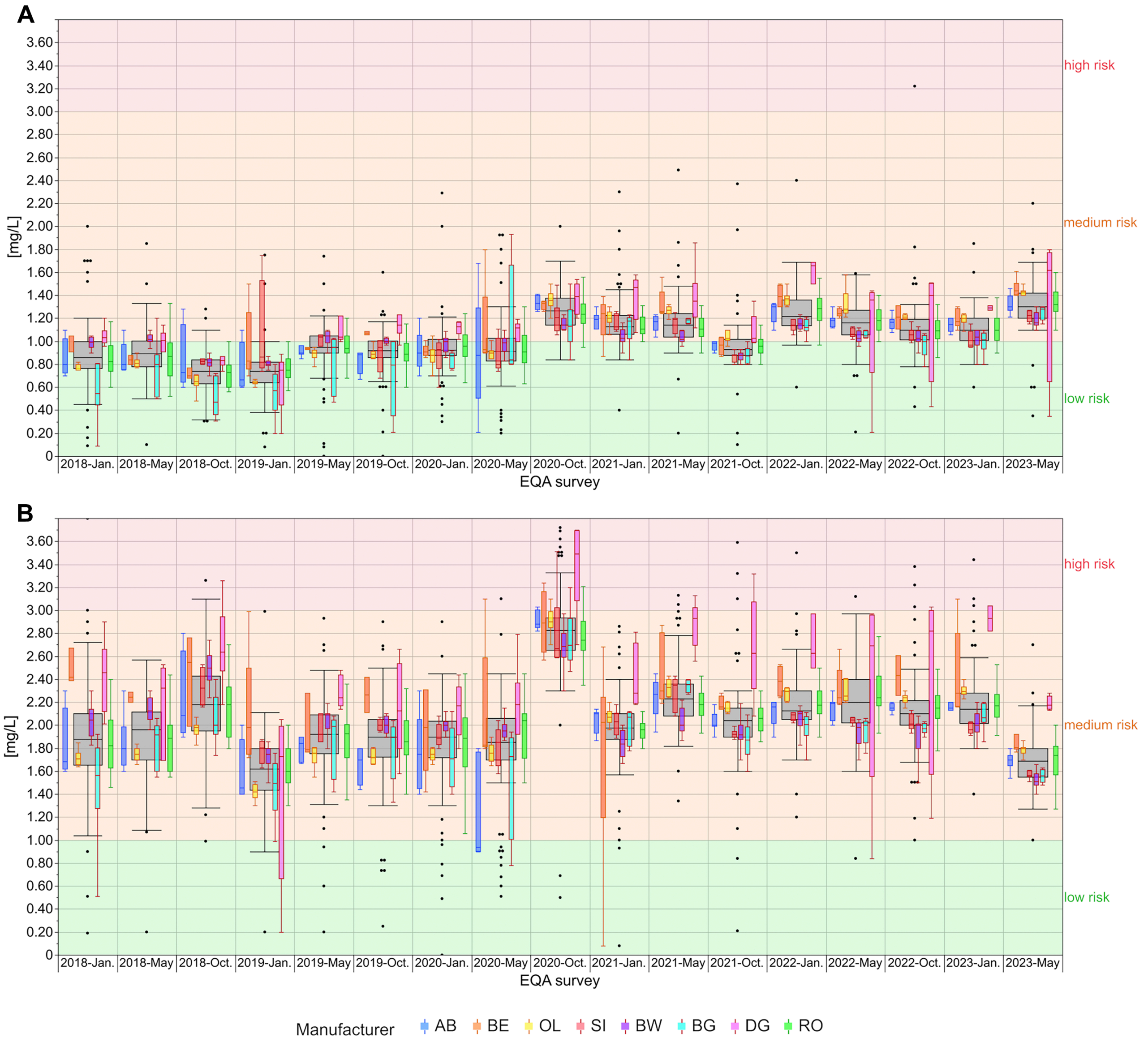


FIGURE 2
Analysis of manufacturer-dependent differences for the detection of hs-CRP in serum from 2018 to 2023 for the low concentration **(A)** and high concentration sample **(B)**. The grey boxes display all results for the respective sample, and the distributions of specific manufacturer-based collectives are illustrated as smaller, colored box plots in overlay with the total results. For all boxes, the whiskers stretch from the 1st quartile - 1.5 ∗ (interquartile range) to the 3rd quartile + 1.5 ∗ (interquartile range). OL is a sub-collective of BE, and BW, BG and DG are sub-collectives of SI, hence the the same outline but different filling. All results below 1 mg/L are considered "low risk for cardiovascular disease" (green area). Results between 1 mg/L and 3 mg/L are considered "medium risk for cardiovascular disease" (orange area) and results above 3 mg/L as "high risk for cardiovascular disease" (red area) (Thomas, 2023).

Manufacturer-dependent differences as well as the scatter of results were found to decrease slightly after October 2020, when the sample material was changed from a serum pool to sera obtained from individual donors. One exception was the DG collective, whose median values increased in comparison to the other collectives, especially for the high concentration samples (Figure 3B). While the median values of the other collectives decreased from over 50% to only 20%, the DG collective exceeded the median of other manufacturer collectives by up to 35% in May 2023 (Figure 3). Promising trends were observed for the scatter of results, as the CVs of most manufacturer collectives, apart from the DG collective, stayed below 10% (Figure 4).

Our results correspond to those of several research groups that have reported similar differences between various assays. Thanabalasingham et al. (2011) observed differences between three assays from SI, labeling them methods one, two, and three. The assay from method one was used by this paper's BG collective and the assay from method three by the BW collective. They observed higher results for the BG assay for hs-CRP concentrations >1 mg/L and higher results for the BW assay for hs-CRP concentrations below

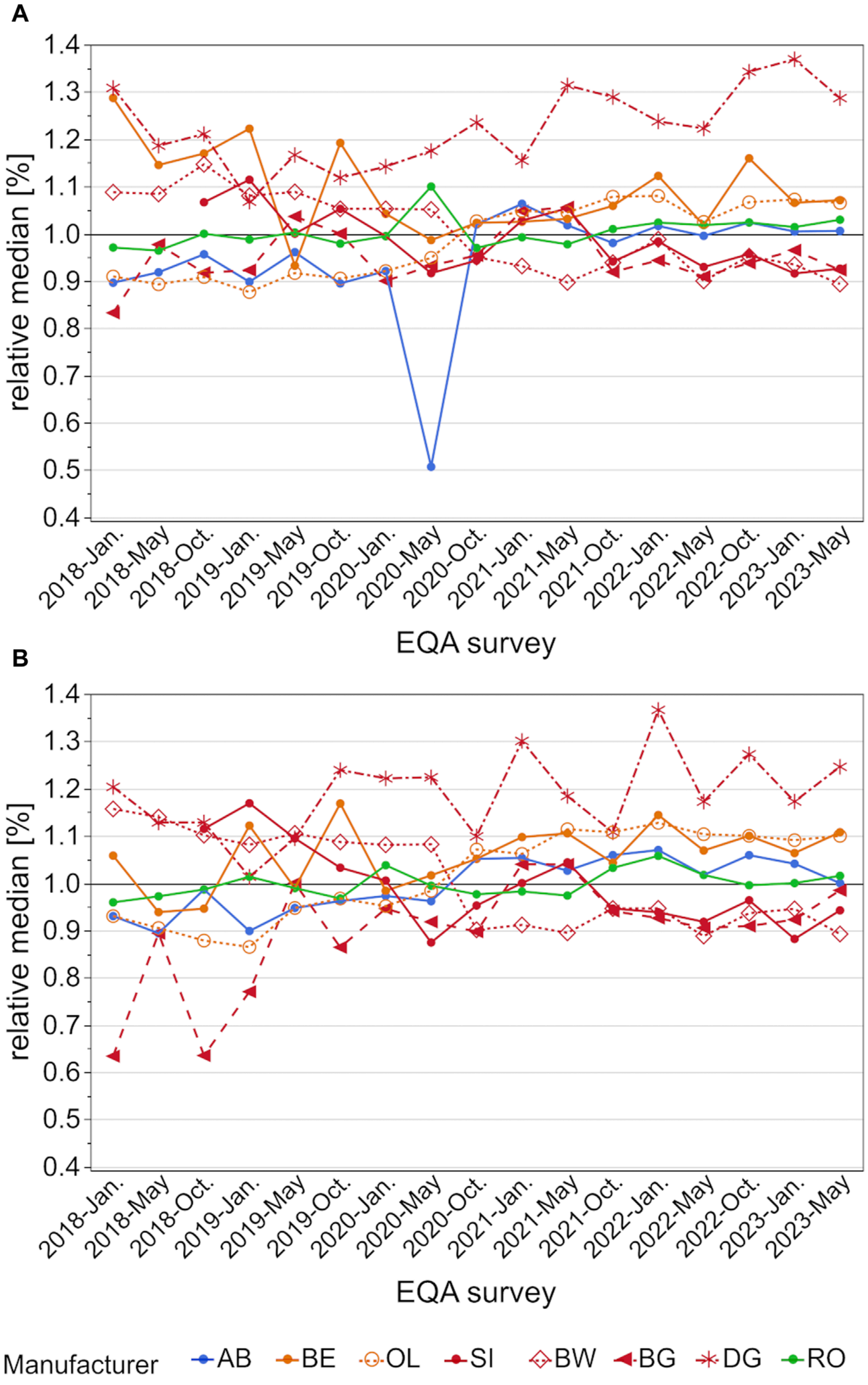


FIGURE 3
Analysis of manufacturer-dependent differences in median values for the detection of hs-CRP in serum from 2018 to 2023 for the low concentration **(A)** and high concentration sample **(B)**. All median values are normalized to the total median of the corresponding EQA scheme. OL is a sub-collective of BE, and BW, BG and DG are sub-collectives of SI, hence the same colors but different pattern.

this threshold. The data from the INSTAND EQA schemes showed that in 2018 the BW collective had up to 50% higher hs-CRP median values than the BG collective, regardless of sample concentration. These differences have nearly vanished since October 2021 and now these two collectives align quite well (Figure 3), possibly due to a re-calibration of the tests.

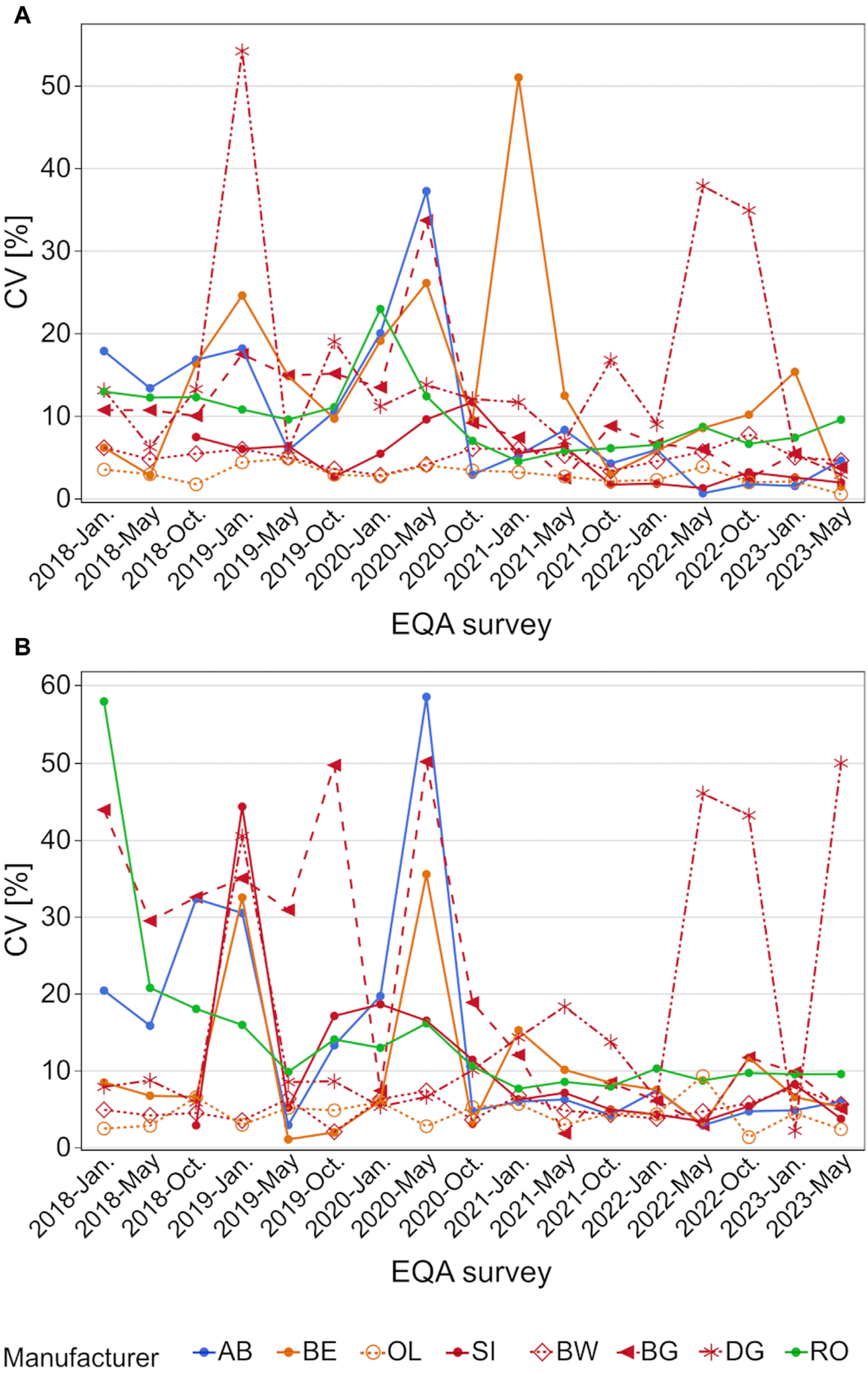


FIGURE 4
Analysis of manufacturer-dependent differences in CV for the detection of hs-CRP in serum from 2018 to 2023 for the low concentration **(A)** and high concentration sample **(B)**. OL is a sub-collective of BE, and BW, BG and DG are sub-collectives of SI, hence the same colors but different pattern.

For the collectives BE and OL, the manufacturer-dependent differences in median values observed in January 2018 decreased over time and nearly vanished from October 2020 onwards (Figure 3).

An older paper by Roberts et al. reported that the assays from AB and BE showed higher results in serum pools with more than 2 mg/L CRP than the test systems from BW (Roberts et al., 2000). In a follow-up study, the test systems from RO and OL showed comparable or slightly higher values than the system from BW (Roberts et al., 2001). Data from the EQA surveys showed that, while the difference between AB, BW, and OL had nearly vanished in the last 2 years, RO still had slightly higher median values than BW (Figure 3).

The positive trend in the harmonization between the hs-CRP assay manufacturers analyzed in this study began appearing in 2019–2020. At that time, the relative median values started to align until the observed differences were below 20% for the highest and lowest collectives, with the exception of the DG collective. Since the trend started before INSTAND changed the sample material from serum pools to single-donor samples, an influence of pooled samples on the median comparison is unlikely. Nevertheless, the change in sample material could have had an influence on the scatter results since, strikingly, the CVs stayed below 10% more often after the change in sample material (Figure 4). The occasional outliers in CV could be due to small sample sizes, e.g., the DG collective showed the highest CVs in 2022 and 2023 when fewer than ten laboratories participated in each EQA survey (Supplementary Table S2).

Another reason for the good harmonization observed in this paper is the presence of a certified reference material for hs-CRP: ERM DA474/IFCC. The CRP value for this standard was assigned using ERM-DA470 as a calibrant, which is traceable to the WHO International Standard 85/506 (Hanisch et al., 2011). The follow-up standard for ERM-DA470, ERM-DA470k/IFCC, was unsuitable for the certification of CRP due to a roughly 20% loss of CRP in the lyophilized standard preparation when compared to frozen material, as measured by routine immunoassays (Zegers et al., 2010).

Interestingly, SI reassigned calibrator lots for their Advia (BG collective) and Atellica (SI collective) hs-CRP assays from ERM-DA470 to ERM DA474/IFCC as they observed a positive bias of approximately 15% for patient samples and quality control material when compared to ERM DA474/IFCC (Siemens Healthineers, 2020a; Siemens Healthineers, 2020b). In the meantime, several RO hs-CRP assays still state their traceability to the ERM-DA470 standard preparation (Roche Diagnostics GmbH, 2023a; Roche Diagnostics GmbH, 2023b).

These differences in the traceability of calibrators could be one factor in why the BW and the SI collectives showed such a clear drop in relative median results around May and October 2020. While the SI collective rose once again for a short time, both collectives showed almost identical results beginning in October 2021. The RO collective exhibited relative median values that were around 10% higher than those of BW and SI starting in October 2021, but they aligned well in May 2023 for the higher concentration sample (see Figure 3).

Systems from BW [e.g., (de Lemos et al., 2017; Tunstall-Pedoe et al., 2017; Khera et al., 2018; Lee et al., 2018; Zhang et al., 2021)] and RO [e.g., (Petersson et al., 2009; Eugen-Olsen et al., 2010)] were most frequently deployed in several meta-analyses. The observed differences in the INSTAND EQA surveys clearly show that the observed assay variability could have a significant impact on the "real-life" CVD risk classification for patients, despite the relatively good harmonization (Figure 2). For example, for the high concentration samples analyzed in May and October 2022 (Figure 2B), 75% of laboratories using BW reported results of <2 mg/L, while over 75% of RO laboratories reported results of >2 mg/L for the same patient. Therefore, meta-analyses that compile data from different clinical studies should not only be stratified by study population and research question, but also by hs-CRP assay to ensure valid data aggregation and interpretation.

One limitation of this study is that it is not possible to assess whether the changes in median results, especially in the sub-collectives of SI, are due to the recalibration of their calibrators or due to the change in sample material from serum pools to single-donor samples. But since a positive trend for harmonization was observed before the change in sample material, it is unlikely that the effect is solely based on that switch. Furthermore, some of the collectives were quite small, which could bias the CV calculation.

The results from this analysis clearly show the high importance of a well-tailored diagnosis and treatment policy in CVD patients. However, while huge efforts have been made to raise the level of assay harmonization for this marker to the current level, new complications appear on the horizon. At a recent JCTLM workshop, data were presented on newly developed primary pure candidate substances and secondary certified reference materials (CRMs). Furthermore, new reference measurement procedures indicate that clinical samples measured with procedures that are calibrated with the CRMs mentioned might clearly differ from results measured by the immunoassays currently calibrated to the existing standard materials. However, the possible influence of such new CRMs or RMPs of higher order on the measurement of CRP and hs-CRP still requires further assessment (Miller et al., 2023).

5 Conclusion

While the harmonization of hs-CRP assays is quite good, the observed bias in the EQA surveys could still lead to a clinical misclassification in the case of risk stratification for CVDs under real life conditions. Although our data do not provide any insight on the dimension of this risk, it is clear that hs-CRP should not be used as a single marker for risk stratification and longitudinal measurements of the same patient and should always be done in the same device. For a better future harmonization, new developments in reference materials and reference measurement procedures for CRP and hs-CRP need to be carefully observed. Especially without a proper reference measurement procedure it is currently impossible to give any recommendations for or against an assay for the detection of hs-CRP as well as its use in a CVD risk score. But meta-analysis of different risk-score publications should stratify their analysis by assay to account for the observed manufacturer-specific differences observed in this paper.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Ethics statement

This study was conducted in accordance with the Statement of the Central Ethics Commission of Germany on the use of human body materials for medical research purposes (no. 20/02/2003; https://www.zentrale-ethikkommission.de). The donor's informed written consent is available for the single donor samples of the project. A positive vote from the ethics committee of Goethe University Frankfurt (Main) has been obtained for samples from voluntary blood donors.

Author contributions

NW: Conceptualization, Formal Analysis, Writing–original draft, Writing–review and editing, Data curation, Visualization. LV: Writing–review and editing, Conceptualization, Visualization. MK: Writing–review and editing. AK-S: Writing–review and editing. SK: Writing–review and editing, Data curation. SB: Writing–review and editing. KK: Writing–review and editing. K-PH: Supervision, Writing–review and editing, Conceptualization. IS: Conceptualization, Supervision, Writing–review and editing.

Funding

The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2024.1401405/full#supplementary-material

References

Ahmad, A., Lim, L. L., Morieri, M. L., Tam, C. H., Cheng, F., Chikowore, T., et al. (2024). Precision prognostics for cardiovascular disease in Type 2 diabetes: a systematic review and meta-analysis. *Commun. Med. (Lond)* 4 (1), 11. doi:10.1038/s43856-023-00429-z

Australian Chronic Disease Prevention Alliance (2023). *Australian Guideline for assessing and managing cardiovascular disease risk*.

CDC (2021). Health topics – heart disease and heart attack. Available at: https://www.cdc.gov/policy/polaris/healthtopics/heartdisease/index.html.

Charoud-Got, J., Zegers, I., Rzychon, M., and Centre, Joint Research, Materials, Institute for Reference and Measurements (2009). *Certification of C-reactive protein in reference material ERM-DA472/IFCC certified reference materials ERM-DA472/IFCC*. Publications Office.

de Lemos, J. A., Ayers, C. R., Levine, B. D., deFilippi, C. R., Wang, T. J., Hundley, W. G., et al. (2017). Multimodality strategy for cardiovascular risk assessment: performance in 2 population-based cohorts. *Circulation* 135 (22), 2119–2132. doi:10.1161/CIRCULATIONAHA.117.027272

Deutsche Gesellschaft für Allgemeinmedizin und Familienmedizin e.V (2017). *Hausärztliche Risikoberatung zur kardiovaskulären Prävention. S3-Leitlinie, AWMF*.

Eugen-Olsen, J., Andersen, O., Linneberg, A., Ladelund, S., Hansen, T. W., Langkilde, A., et al. (2010). Circulating soluble urokinase plasminogen activator receptor predicts cancer, cardiovascular disease, diabetes and mortality in the general population. *J. Intern Med.* 268 (3), 296–308. doi:10.1111/j.1365-2796.2010.02252.x

European Society of Cardiology (2023). *Price tag on cardiovascular disease in Europe higher than entire EU budget*.

Hanisch, K., Zegers, I., Schimmel, H., Boulo, S., Emons, H., Trapmann, S., et al. (2011). *The certification of the mass concentration of C-reactive protein in human serum – certified reference material ERM®-DA474/IFCC. Publications Office.*

Heidemann, C., Scheidt-Nave, C., Beyera, A.-K., Baumert, J., Thamm, R., Maier, B., et al. (2021). Gesundheitliche Lage von Erwachsenen in Deutschland-Ergebnisse zu ausgewählten Indikatoren der Studie GEDA 2019/2020-EHIS. *J. Health Monit.* 6 (3), 3–27.

International Organization for Standardization [ISO] (2023). *Konformitätsbewertung–Allgemeine Anforderungen an die Kompetenz von Anbietern von Eignungsprüfungen (ISO/IEC 17043:2023); Deutsche und Englische Fassung EN ISO/IEC 17043:2023. DDIfNe V.*

Khera, A., Budoff, M. J., O'Donnell, C. J., Ayers, C. A., Locke, J., de Lemos, J. A., et al. (2018). Astronaut cardiovascular Health and risk modification (Astro-CHARM) coronary calcium atherosclerotic cardiovascular disease risk calculator. *Circulation* 138 (17), 1819–1827. doi:10.1161/CIRCULATIONAHA.118.033505

Lee, D. Y., Rhee, E. J., Chang, Y., Sohn, C. I., Shin, H. C., Ryu, S., et al. (2018). Impact of systemic inflammation on the relationship between insulin resistance and all-cause and cancer-related mortality. *Metabolism* 81, 52–62. doi:10.1016/j.metabol.2017.11.014

Lee, H. S., and Lee, J. H. (2023). Early elevation of high-sensitivity C-reactive protein as a predictor for cardiovascular disease incidence and all-cause mortality: a landmark analysis. *Sci. Rep.* 13 (1), 14118. doi:10.1038/s41598-023-41081-w

Li, Y., Zhong, X., Cheng, G., Zhao, C., Zhang, L., Hong, Y., et al. (2017). Hs-CRP and all-cause, cardiovascular, and cancer mortality risk: a meta-analysis. *Atherosclerosis* 259, 75–82. doi:10.1016/j.atherosclerosis.2017.02.003

Libby, P., Ridker, P. M., and Maseri, A. (2002). Inflammation and atherosclerosis. *Circulation* 105 (9), 1135–1143. doi:10.1161/hc0902.104353

Lin, J. S., Evans, C. V., Johnson, E., Redmond, N., Coppola, E. L., and Smith, N. (2018). Nontraditional risk factors in cardiovascular disease risk assessment: updated evidence report and systematic review for the US preventive Services Task Force. *Jama* 320 (3), 281–297. doi:10.1001/jama.2018.4242

Lobo, S. M. (2012). Sequential C-reactive protein measurements in patients with serious infections: does it help? *Crit. Care* 16 (3), 130. doi:10.1186/CC11347

Miller, W. G., Panteghini, M., and Wielgosz, R. (2023). Implementing metrological traceability of C-reactive protein measurements: consensus summary from the joint committee for traceability in laboratory medicine workshop. *Clin. Chem. Lab. Med.* 61 (9), 1558–1560. doi:10.1515/cclm-2023-0498

Müller, I., Besier, S., Hintereder, G., Brade, V., and Hunfeld, K. (2009). Zur Qualität der bakteriologischen Infektionsserologie in Deutschland: eine Metaanalyse der infektionsserologischen Ringversuche des Jahres 2006—Beitrag der Qualitätssicherungskommission der DGHM. *GMS Z Forder Qualitatssich Med. Lab.* 1, 1–21.

Musunuru, K., Kral, B. G., Blumenthal, R. S., Fuster, V., Campbell, C. Y., Gluckman, T. J., et al. (2008). The use of high-sensitivity assays for C-reactive protein in clinical practice. *Nat. Clin. Pract. Cardiovasc Med.* 5 (10), 621–635. doi:10.1038/ncpcardio1322

Oude Wolcherink, M. J., Behr, C. M., Pouwels, X., Doggen, C. J. M., and Koffijberg, H. (2023). Health economic research assessing the value of early detection of cardiovascular disease: a systematic review. *Pharmacoeconomics* 41 (10), 1183–1203. doi:10.1007/s40273-023-01287-2

Pepys, M. B., and Hirschfield, G. M. (2003). C-reactive protein: a critical update. *J. Clin. investigation* 111 (12), 1805–1812. doi:10.1172/JCI18921

Petersson, U., Ostgren, C. J., Brudin, L., and Nilsson, P. M. (2009). A consultation-based method is equal to SCORE and an extensive laboratory-based method in predicting risk of future cardiovascular disease. *Eur. J. Cardiovasc Prev. Rehabil.* 16 (5), 536–540. doi:10.1097/HJR.0b013e32832b1833

Piepoli, M. F., Hoes, A. W., Agewall, S., Albus, C., Brotons, C., Catapano, A. L., et al. (2016). 2016 European guidelines on cardiovascular disease prevention in clinical practice: the sixth joint Task Force of the European society of Cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of 10 societies and by invited experts)Developed with the special contribution of the European association for cardiovascular prevention and rehabilitation (EACPR). *Eur. Heart J.* 37 (29), 2315–2381. doi:10.1093/eurheartj/ehw106

Póvoa, P. (2002). C-reactive protein: a valuable marker of sepsis. *Intensive Care Med.* 28 (3), 235–243. doi:10.1007/s00134-002-1209-6

Póvoa, P., Almeida, E., Moreira, P., Fernandes, A., Mealha, R., Aragão, A., et al. (1998). C-reactive protein as an indicator of sepsis. *Intensive Care Med.* 24 (10), 1052–1056. doi:10.1007/s001340050715

Ridker, P. M., Danielson, E., Fonseca, F. A., Genest, J., Gotto, A. M., Jr., Kastelein, J. J., et al. (2008). Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *N. Engl. J. Med.* 359 (21), 2195–2207. doi:10.1056/NEJMoa0807646

Ridker, P. M., Everett, B. M., Thuren, T., MacFadyen, J. G., Chang, W. H., Ballantyne, C., et al. (2017). Antiinflammatory therapy with canakinumab for atherosclerotic disease. *N. Engl. J. Med.* 377 (12), 1119–1131. doi:10.1056/nejmoa1707914

Roberts, W. L., Moulton, L., Law, T. C., Farrow, G., Cooper-Anderson, M., Savory, J., et al. (2001). Evaluation of nine automated high-sensitivity C-reactive protein methods: implications for clinical and epidemiological applications. Part 2. *Clin. Chem.* 47 (3), 418–425. doi:10.1093/clinchem/47.3.418

Roberts, W. L., Sedrick, R., Moulton, L., Spencer, A., and Rifai, N. (2000). Evaluation of four automated high-sensitivity C-reactive protein methods: implications for clinical and epidemiological applications. *Clin. Chem.* 46 (4), 461–468. doi:10.1093/clinchem/46.4.461

Roche Diagnostics GmbH (2023a). "Cardiac C-reactive protein (latex) high sensitive (cobas c 111)." V 9.0 English. Available at: https://elabdoc-prod.roche.com/eLD/api/downloads/fbc6a02f-8a83-ee11-2291-005056a71a5d?countryIsoCode=pi.

Roche Diagnostics GmbH (2023b). "Cardiac C-reactive protein (latex) high sensitive (cobas c 303, cobas c 503)." V 9.0 English. Available at: https://elabdoc-prod.roche.com/eLD/api/downloads/fbc6a02f-8a83-ee11-2291-005056a71a5d?countryIsoCode=pi.

Romero-Cabrera, J. L., Ankeny, J., Fernández-Montero, A., Kales, S. N., and Smith, D. L. (2022). A systematic review and meta-analysis of advanced biomarkers for predicting incident cardiovascular disease among asymptomatic middle-aged adults. *Int. J. Mol. Sci.* 23 (21), 13540. doi:10.3390/ijms232113540

Shah, T., Casas, J. P., Cooper, J. A., Tzoulaki, I., Sofat, R., McCormack, V., et al. (2009). Critical appraisal of CRP measurement for the prediction of coronary heart disease events: new data and systematic review of 31 prospective cohorts. *Int. J. Epidemiol.* 38 (1), 217–231. doi:10.1093/ije/dyn217

Siemens Healthineers (2020a). *Reassignment of the ADVIA®chemistry CardioPhase high sensitivity C-reactive protein (hsCRP) calibrator lots 484707 and 516407. Tarrytown, NY 10591.*

Siemens Healthineers (2020b). *Reassignment of the Atellica CH high sensitivity C-reactive protein (hsCRP) calibrator lots 484721 and 516427. Tarrytown, NY 10591.*

Statistisches Bundesamt (2022). *Krankheitskosten pro Kopf gleichen sich zwischen Männern und Frauen weiter an.*

Stevenson, E., Walsh, C., and Thomas, S. (2023). EQA: there's not a glitch in the matrix. Investigation of CRP bias on the Roche Cobas c701. *Ann. Clin. Biochem.* 60 (5), 349–352. doi:10.1177/00045632231169151

Thanabalasingham, G., Shah, N., Vaxillaire, M., Hansen, T., Tuomi, T., Gašperíková, D., et al. (2011). A large multi-centre European study validates high-sensitivity C-reactive protein (hsCRP) as a clinical biomarker for the diagnosis of diabetes subtypes. *Diabetologia* 54 (11), 2801–2810. doi:10.1007/s00125-011-2261-y

Thomas, L. (2023). *Atherosclerosis - development of atherosclerosis*. Clinical laboratory diagnostics.

Tunstall-Pedoe, H., Peters, S. A. E., Woodward, M., Struthers, A. D., and Belch, J. J. F. (2017). Twenty-year predictors of peripheral arterial disease compared with coronary heart disease in the scottish heart Health extended cohort (SHHEC). *J. Am. Heart Assoc.* 6 (9), e005967. doi:10.1161/JAHA.117.005967

U.S. Preventive Services Task Force (2009). Using nontraditional risk factors in coronary heart disease risk assessment: U.S. Preventive Services Task Force recommendation statement. *Ann. Intern Med.* 151 (7), 474–482. doi:10.7326/0003-4819-151-7-200910060-00008

Visseren, F. L. J., Mach, F., Smulders, Y. M., Carballo, D., Koskinas, K. C., Bäck, M., et al. (2022). 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice: developed by the Task Force for cardiovascular disease prevention in clinical practice with representatives of the European Society of Cardiology and 12 medical societies with the special contribution of the European Association of Preventive Cardiology (EAPC). *Rev. Esp. Cardiol. Engl. Ed.* 75 (5), 429. doi:10.1016/j.rec.2022.04.003

WHO (2023). Noncommunicable diseases. Available at: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.

Wojtalewicz, N., Schellenberg, I., and Hunfeld, K. P. (2019). Evaluation of INSTAND e.V.'s external quality assessment for C-reactive protein and procalcitonin. *PLoS One* 14 (8), e0221426.

Zegers, I., Schreiber, W., Linstead, S., Lammers, M., McCusker, M., Muñoz, A., et al. (2010). Development and preparation of a new serum protein reference material: feasibility studies and processing. *Clin. Chem. Lab. Med.* 48 (6), 805–813. doi:10.1515/CCLM.2010.166

Zhang, W., Speiser, J. L., Ye, F., Tsai, M. Y., Cainzos-Achirica, M., Nasir, K., et al. (2021). High-Sensitivity C-reactive protein modifies the cardiovascular risk of lipoprotein(a): multi-ethnic study of atherosclerosis. *J. Am. Coll. Cardiol.* 78 (11), 1083–1094. doi:10.1016/j.jacc.2021.07.016

TYPE Original Research
PUBLISHED 26 March 2024
DOI 10.3389/fimmu.2024.1368399

OPEN ACCESS

EDITED BY
Uday Kishore,
United Arab Emirates University,
United Arab Emirates

REVIEWED BY
Marina Noris,
Mario Negri Institute for Pharmacological Research (IRCCS), Italy
Matteo Stravalaci,
IRCCS Ca 'Granda Foundation Maggiore Policlinico Hospital, Italy

*CORRESPONDENCE
Zoltán Prohászka
✉ prohaszka.zoltan@semmelweis.hu

RECEIVED 23 January 2024
ACCEPTED 07 March 2024
PUBLISHED 26 March 2024

CITATION
Kirschfink M, Frazer-Abel A, Balogh E, Goseberg S, Weiss N and Prohászka Z (2024) External quality assurance program for diagnostic complement laboratories: evaluation of the results of the past seven years. *Front. Immunol.* 15:1368399. doi: 10.3389/fimmu.2024.1368399

External quality assurance program for diagnostic complement laboratories: evaluation of the results of the past seven years

Michael Kirschfink[1,2], Ashley Frazer-Abel[3], Emese Balogh[4], Sabine Goseberg[2], Nathalie Weiss[2] and Zoltán Prohászka[2,5]*

[1]Institute of Immunology, University of Heidelberg, Heidelberg, Germany, [2]Instand eV, Düsseldorf, Germany, [3]ExseraBioLabs, University of Colorado, Aurora, CO, United States, [4]Department of Pharmaceutics, Semmelweis University, Budapest, Hungary, [5]Department of Internal Medicine and Hematology, Füst György Complement Diagnostic Laboratory, Semmelweis University, Budapest, Hungary

Introduction: The complement external quality assurance (EQA) program was first organized in 2010 by a group of researchers working in diagnostic complement laboratories. Starting in 2016, INSTAND e.V., a German, non-profit interdisciplinary scientific medical society dedicated to providing expert EQA programs for medical laboratories, started organizing the EQAs for complement diagnostic laboratories together with the same group of experienced scientists and doctors who also work as EQA experts. The aim of the current work is to provide descriptive analysis of the past seven years' complement EQA results and evaluate timeline changes in proficiency testing.

Methods: Each year, in March and October, blinded samples (normal, pathological) were sent to the participating diagnostic laboratories, where complement parameters were evaluated exactly as in daily routine samples. Since no reference method/target values exist for these parameters, and participants used different units for measurement, the reported results were compared to the stable mean (Algorithm A) of the participants using the same method/measurement units. A reported result was qualified as "passed" if it fell into the 30-50% evaluation/target range around the mean of reported results (depending on the given parameter).

Results: While the number of participating laboratories has increased in the past years (from around 120 to 347), the number of complement laboratories providing multiple determinations remained mostly unchanged (around 30 worldwide). C3, C4, C1-inhibitor antigen and activity determinations provided the best proficiency results, with >90% passing quotas in the past years, independent of the applied method. Determination of the functional activity of the three activation pathways was good in general, but results showed large variance, especially with the pathological samples. Complement factor C1q and regulators FH and FI are determined by only a few laboratories, with variable outcomes (in general in the 85-90% pass range). Activation products sC5b-9 and Bb were determined in 30 and 10 laboratories, respectively, with typical passing quotas in the 70-90% range, without a clear tendency over the past years.

Conclusion: With these accumulated data from the past seven years, it is now possible to assess sample-, method-, and evaluation related aspects to further improve proficiency testing and protocolize diagnostic complement determinations.

KEYWORDS

external quality assessment schemes, proficiency testing, complement pathways, functional analysis, complement activation products, autoantibody, C3 nephritic factor

Introduction

As a major part of innate immune system, complement is not only essential to fight invading pathogens but also plays a key role in immune surveillance, homeostasis and repair (1–3). More than 50 soluble and cell-bound proteins serve either as danger sensing molecules for invading pathogens, apoptotic and necrotic cells and immune complexes (e.g. C1q, mannose-binding lectin/MBL, ficolins, properdin, proteins of the Factor H family). These molecules act within an enzymatic cascade and provide a very effective regulation of immunity via receptors on the surface of multiple immune-, and tissue cells (for review see (4, 5). Complement proteins in the circulation are primarily synthesized in the liver and by monocytes, but are also constitutively expressed and secreted by many other cell types in different tissues into the microenvironment (6, 7).

Upon activation via the classical (CP), lectin (LP) or alternative pathway (AP), the central components C3 and C5 are cleaved, which results in the opsonization of pathogens and debris with C3b and iC3b, the recruitment of inflammatory cells via the anaphylatoxins C3a and C5a, and finally in the formation of the membrane attack complex, C5b-9 (reviewed in (8–10). Under physiological conditions, the system is tightly regulated by proteins in the fluid phase (CP: C1-inhibitor, C4 binding protein, Factor I; AP: Factor H, Factor I; LP: C1-inhibitor, terminal pathway: clusterin, vitronectin/S-protein). Membrane-bound inhibitors protect each individual cell in the circulation and solid tissue (CD35/CR1, CD46/MCP CD55/DAF and CD59) to prevent unwanted activation (11, 12). For schematic illustration of complement pathways and regulation see Figure 1 in (13).

While the primary action of complement is well described for plasma and body fluids in the extravascular space, more recent studies suggest a possible role also inside the cell (14). Multiple interactions exist between the coagulation, fibrinolytic and complement systems where enzymes can cleave and activate one another, and regulators are shared between cascades (15). This provides a good explanation why many complement-driven diseases (e.g. PNH, aHUS, CHAPLE syndrome) express thrombosis as a hallmark of clinical manifestation (16, 17).

Complement deficiencies comprise about 5-10%, according to different registries of all primary immunodeficiencies with a combined genetic prevalence of 0.03% in the general population. Probably clinically more relevant are consequences of complement overactivation leading to numerous inflammatory and autoimmune diseases (18–20).

In the last decades great progress has been made in complement analysis to not only understand its physiology but also to better define disease development, severity, and response to therapy (21). This has been further accelerated by the introduction of complement-targeting drugs, which has led to a significant increase of interest by clinicians (13, 22).

A comprehensive laboratory analysis of the complement system should start with the assessment of the total activity of the classical and alternative pathway either by functional ELISA or by hemolytic or liposome-based assays (23). These global tests provide information about the integrity of the entire complement cascade. A missing or greatly reduced activity indicates a primary complement deficiency but may also be due to a secondary deficiency caused by decreased synthesis, increased consumption, or protein loss. Deficient or dysfunctional proteins of the affected pathway are identified by single protein (ELISA, radial immunodiffusion, immunoelectrophoretic or nephelometry/turbidimetry) or functional tests (24, 25).

Since most of the complement components are acute-phase proteins with a higher rate of synthesis in inflammation, in acute-phase reaction individual components are often left within the normal range despite ongoing consumption. Only the analysis of complement activation products allows one to distinguish with enough sensitivity complement deficiency from pathologically increased complement activation and consumption *in vivo* (26). Complement activation products may be either split fragments after enzymatic cleavage of certain components, e.g., C4 (C4a, C4b/c, C4d), C3 (C3a, C3b, iC3b, C3c,C3dg, C3d), FB (Ba, Bb), and C5 (C5a), or protein complexes where activated components are bound to their respective regulators, like C1rs–C1-INH, the properdin-containing alternative pathway convertase C3bBbP, and sC5b-9 (soluble terminal complement complex, also known as soluble membrane attack complex sMAC, or terminal complement complex TCC). Quantification can be done as traditional ELISA, upon binding to high-capacity immunosorbent with subsequent elution, or to microbeads applied in multiplex flow cytometric technology (see below). Those neoepitope-specific antibodies are

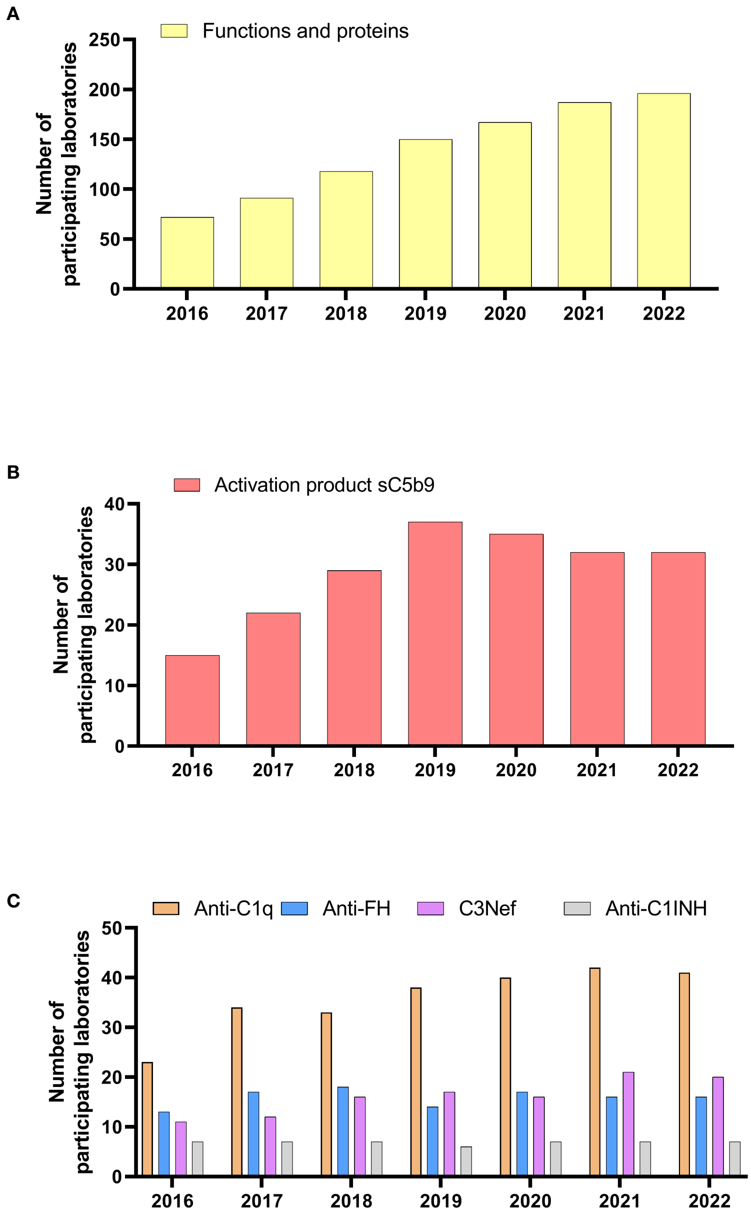


FIGURE 1
Numbers of participating laboratories in the different complement EQA programs in the past years. **(A)** EQA246, **(B)** EQA247, **(C)** EQA245, EQA248, EQA249 and EQA250. For EQA245 and EQA246 numbers of laboratories with at least 1 participation in the indicated year are given.

also valuable to detect *in situ* complement activation applying immunohistochemistry.

Sensitive and quantitative multiplex analysis tests are currently developed to simultaneously assess multiple complement proteins and activation products but have not yet been applied to routine complement analysis (27).

Importantly, routine laboratory analysis of complement abnormalities also involves the measurement of clinically relevant inhibitory or activating autoantibodies targeting individual complement components, regulators, or convertases such as C1 inhibitor, C1q, Factor H, and C3 nephritic factor. These autoantibodies have been demonstrated to be useful as diagnostic or prognostic markers as well as for monitoring therapeutic responses (28).

Preanalytical considerations are important determinants of quality of results in the diagnostic complement laboratory (29). As outlined in a recent review by Brandwijk et al. about 50% of all investigated studies failed to use the right sample type or technique (30).

Since many complement proteins are heat labile precise preanalytical sample handling is mandatory for accurate and conclusive laboratory complement diagnostics. Correct collection and processing of all body fluid samples for complement analysis is essential to avoid artificial *ex vivo* complement activation. Without inhibition, physiological and pathological complement activation continues *ex vivo* obscuring the actual complement activation status and preventing meaningful data interpretation.

Serum is the appropriate sample to measure complement activity, components, regulators, and autoantibodies. It should be separated by centrifugation after full clotting and samples should be used immediately or can be stored at -70 °C for longer times. Since multiple serine proteases from other cascade systems can cleave complement components it is strongly recommended to use EDTA-plasma for analyzing complement activation products. Heparin and citrate-based anticoagulants are less useful (31).

For most complement activation products, EDTA-plasma is stable for up to 4 hours at room temperature (32) but should better be kept on ice or in a refrigerator if analyzed on the same day. For later processing, the sample should be aliquoted, frozen, and stored at -70 °C. Frozen samples should be thawed at room temperature or on ice, but not in a water bath at 37 °C. Repeated freezing and thawing of aliquots should be avoided. Frozen samples must be shipped on dry ice by courier if transport is necessary. Samples must be collected prior to plasma infusion or plasma exchange, or before any kind of immune therapies causing complement mediated cytolysis (for example anti-CD20 or anti-thymocyte globulin therapies) to determine the initial disease-related complement status.

In urine, the measurement of complement activation products can be affected by high amounts of urea and urine proteases. Since activation products in proteinuria may appear as a consequence of extrarenal (artificial) rather than intrarenal complement activation, the addition of protease inhibitors is required (33).

Complement proteins can also be analyzed in bronchoalveolar lavage (34), cerebrospinal (35) or synovial fluids (36) as well as tears and aqueous/vitreous humor (37) which may better reflect a local complement activation.

Finally, correct interpretation requires validated reference intervals. Here it should be emphasized that the reference intervals for several components are age-related, especially when analyzing samples from infants this must be taken into account (38–40).

Following the increased attention for complement analysis over the last 2 decades and a need to improve its consistency and quality the Sub-Committee for the Standardization and Quality Assessment of Complement Measurements was established and formally recognized by the IUIS (https://iuis.org/committees/qas/subcommittee-for-the-standardization-and-quality-assessment-of-complement-measurements/). Since 2010, 20 rounds of external quality assessment (EQA), now covering up to 20 parameters (function, proteins, activation products and autoantibodies) have been completed. The aim of the current work is to provide descriptive analysis of the past seven years' complement EQA results and evaluate timeline changes in proficiency testing.

Methods

Sample materials – properties and preparation

In each EQA survey, two samples with normal or pathological concentrations/activities of complement parameters were distributed to the participating laboratories for quantitative or qualitative analysis (Supplementary Table 1). The samples were obtained from either voluntary blood donors or from patients. The samples tested negative for HIV, HBV, and HCV. No stabilizing additives were added (except for EQA247 where EDTA-anticoagulated plasma sample is provided). Samples were lyophilized due to stability reasons for EQA schemes EQA246 and EQA247, and since 2022 also for EQA248. Before 2018 the samples were lyophilized by a commercial provider (in.vent Diagnostica GmbH, Henningsdorf, Germany). Afterwards, the process was done in the Department of Pharmaceutics at Semmelweis University: 1.0 mL in case of EQA246, otherwise 0.3 mL were aliquoted in polypropylene cryo tubes (1.0 and 0.5 mL; Sarstedt, Nümbrecht, Germany). Before 2022, sample volumes differed based on the survey (Supplementary Table 1). The freeze-drying was performed in a one-chamber type equipment (ScanvacCoolsafe 110-04, LaboGene^{TM}, Lynge, Denmark) containing a two shelf sample holder. The process was controlled by a computer program, the temperature and pressure values were recorded continuously. The temperature of the drying chamber was between -97 °C and -95 °C for successful condensation. The samples were previously frozen and kept at -70 °C until the start of lyophilization. The lyophilization started at -40 °C for 1 hour, then temperature of the shelf was increased to the range between 0 °C and 30 °C for 18 hours during the primary drying under 0.02-0.03 hPa vacuum. The secondary drying was performed at 40 °C shelf temperature for 3 hours, where the sample temperature did not exceed 10 °C. The entire lyophilization process took 22 hours. The stability and homogeneity of EQA samples were confirmed according to DIN EN ISO/IEC 17043:2023. All samples were stored at -18 °C until dispatched to participants at ambient temperature.

Ethics statement

The patient's informed written consent is available for the project. A positive vote from the Scientific and Research Committee of the Medical Research Council of Hungary has been obtained. The study was conducted according to the declaration of Helsinki.

EQA procedure

The INSTAND e.V. EQA schemes for analyzing complement parameters are offered worldwide once or twice per year, depending on the scheme. EQA schemes, that are only provided once per year are shipped in October (O) and EQA schemes, that are provided twice a year, are shipped in March (M) and October (O). For detailed information on the different parameters included in each EQA scheme, see Supplementary Table 1. Participating laboratories provide their laboratory results and information on the respective method and reagent provider via the platform RV-Online (http://rv-online.instandev.de). For the evaluation of quantitative results, the consensus value (stable mean) of all participants, calculated using algorithm A, was used (41). Evaluation area around this consensus value depended on the parameter. Detailed information can be found in Supplementary Table 1. With respect to the qualitative results, the participants had to indicate whether the samples were positive, borderline, or negative. The evaluation of qualitative results is based on prior expert evaluation in the laboratory providing the test material.

Qualitative EQA data can be found in Supplementary Table 2 and quantitative EQA data in Supplementary Table 3.

When a manufacturer-dependent variance was observed, collectives were formed and evaluated separately.

Data analysis and statistics

Data are presented as numbers (%) of participants and mean (with SD) of passing quotas (for samples, or for groups). Sample performance rates (passing quota,%) were calculated in the following way: number of laboratories providing results in the target range for a given sample, divided by the number of all laboratories providing results for that given sample. Total rate (passing quota of the group) was calculated in the following way: number of laboratories providing results in the target range for both samples, divided by the number of all laboratories presenting results for both samples.

Statistica 13.5 and GraphPad Prism 9 softwares were used for statistical analysis and data presentation.

Results

The complement EQA program, participation

The EQA program of diagnostic complement laboratories comprises six schemes: EQA246 (ten parameters) for complement function, components, and regulators, EQA247 (four parameters) for complement activation products, EQA245 for IgG anti-C1q autoantibody, EQA248 for C3-nephritic factor, EQA249 for IgG anti-FH autoantibody and EQA250 (three parameters) for anti-C1-inhibitor autoantibodies (see Supplementary Table 1). For EQA246, participation markedly increased by 170% in the past seven years (Figure 1A), with almost three times more laboratories participating in 2022, as compared to 2016. The highest increase in participation was observed for C3, C4, C1q, C1-inhibitor concentration (C1INH : Ag) and C1INH function (C1-INHF) (Figure 2). This contrasts with EQA247, where only the terminal pathway activation marker sC5b-9 was measured in at least eight laboratories per year (Figure 1B). Participation peaked in 2019 with a small decline afterwards. Participation for the complement related autoantibodies show great variance (Figure 1C). For anti-C1q and C3Nef there is a clear increase (by 78% and 81%, from 2016 to 2022, respectively). Participation for anti-FH and anti-C1INH remained unchanged in the past years.

For each of the six complement EQA schemes, two blinded samples were offered to the participants: one with normal/negative, and a second with pathological/positive parameter level. Since no reference method or target values exist for these parameters, and participants used different units for their data, the reported results were compared to the stable mean of the participants using the same method/measurement units, if there were at least eight participants in that given subgroup. A reported result was qualified as "passed", if it fell within the 30-50% range around the stable mean (depending on the given parameter). For autoantibody determinations the participants had to report qualitative results using their own cut-off values. In the next paragraphs, EQA performance results are reported as passing quota, indicating the percentage of participants having "passed" in a given EQA scheme. Note, that passing quota was not calculated for subgroups with fewer than eight participants.

Performance of the participants for complement function and proteins

Figure 3 shows mean (with SD) passing quota of 2017-2022 results for EQA246 parameters, separately for the normal (pool of healthy blood donor's serum samples) and the pathological (mixture of normal and heated serum sample of healthy donors) samples. Best performing tests were those for C3 and C4 (not presented, passing quota all the time above 90%, mainly measured by nephelometry or turbidimetry). For C1-inhibitor antigen (measured mainly by nephelometry or turbidimetry) and -function a comparable good performance was observed. The passing quota only occasionally fell below 90%. For C1-inhibitor function approximately two-thirds of the participants used a chromogenic assay (manual or automated), whereas one third used a functional ELISA, both methods and all platforms providing consistently good outcomes (Figure 4).

For determination of complement activity (CP, AP and LP), passing quota on average was higher for normal than for the pathological samples (Figure 3), and this observation is almost constantly present across the years (Figure 4, for CP and AP). CP activity was measured about equally often by each of three methods, based on sheep red blood cell (SRBC) hemolysis, on liposome lysis, or on functional ELISA (detection of C9-neoepitope). Figure 4 shows passing quota separately for these three methods for the last six years (12 EQA surveys). We observed a high variance (70-80% to 100%), with a slightly better performance for the hemolytic assay. Results for AP activity determination were similar in the range of 70% to 100%, without a clear trend or difference in the data over the years or method subgroups (hemolytic or functional ELISA based method).

Several efforts were made in the past to harmonize functional testing in the complement laboratories, either by assay calibration (test sample compared to a normal pool assigned as 100%) or scaling (percentage). Furthermore, the various functional assays yielded nearly similar performance results for C1-inhibitor function, CP and AP activity (Figures 3, 4). However, despite these efforts the raw data from the past years remain divergent between the different functional methods (Figure 5) which indicates that those measurements and results are not interchangeable.

Performance of the participants for complement activation products

Four parameters (C3a, C3d, Bb and sC5b-9) are included in scheme EQA247 for complement activation product, in which two

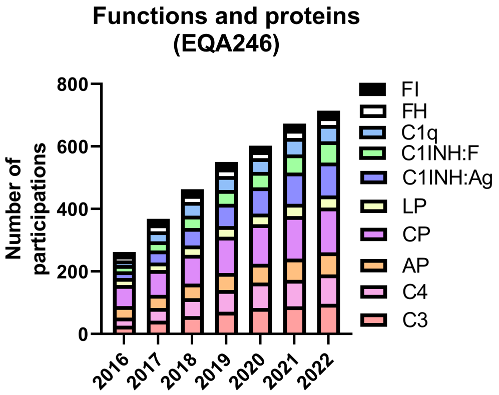


FIGURE 2
Cumulative number of participations (split by various parameters) from laboratories submitting at least one result in the indicated year. Note, that EQA246 was offered twice a year, and the majority of the participants submitted results twice.

blinded, lyophilized samples (0.5 mL normal EDTA plasma, 0.5 mL EDTA plasma spiked with serum of the same donor) are sent to the participants.

Figure 6 illustrates primary measurement results for sC5b-9, a marker measured largely by the same sandwich ELISA (Quidel A029 assay). Approximately 20-25% of the participants with home-based methods could not be analyzed due to differences in assay calibration/ scale. Despite a good correlation over the past six years, every year there were outliers, especially in the upper range of the measurement scale with passing quota for both samples in the 65%-80% range.

Analysis for Bb was less informative, since participant numbers in the past six years varied from eight in 2017, to eleven, ten, eight, eight, seven, in the following years. During these years performance (passing quota) was 75%, 73%, 90%, 89%, 20% and 70%, respectively. Analysis for C3a and C3d was not feasible in the past years since participation remained constantly below eight laboratories.

Performance of the participants for complement autoantibodies

EQA scheme EQA245 for determination of IgG anti-C1q autoantibodies was conducted twice a year, with a total cumulative participation of 69 laboratories. Approximately half of the participants used the same assay (provided by Orgentec), making it possible to analyze the performance separately from participants who used in-house assays or those from different providers (Tables 1A, B). In the Orgentec group, only 4/30 participants performed <80%, whereas in the in-house/INOVA group <80% performance occurred in 11/39 laboratories. It has to be mentioned that nearly half of the participants (14/30) in the Orgentec group, and 12/39 in the in-house/INOVA method group participated regularly, and performance among these frequent attendees was almost exclusively above 80% (bold facing in Tables 1A, B) in both groups.

Interest for C3-nephritic factor (EQA248) increased over time with 28 laboratories participating at least once in the past seven years. Among these 28 laboratories 17 participated at least four times, but performance was above 80% for only 5/17 of the participants (bold facing in Table 2). For the remaining frequently participating laboratories performance was below 80%, and for laboratories with less than four participations proficiency was between 0% and 100%. No clear improvement or change in performance was noted in the past seven years. It has to be noted that participants used a large variety of methods for C3Nef determination. Due to the low number, even in the subgroup using the most frequently applied sheep red blood cell hemolysis based method it was impossible to compare the performance in subgroups discriminating for the applied method.

In contrast, interest for anti-FH didn't change in the past years. 32 laboratories participated at least once, and performance was constantly above 80% (except for 2019) even four times above 90%. From fifteen frequently reporting laboratories, thirteen consistently performed well (bold facing in Table 3). The remaining seventeen

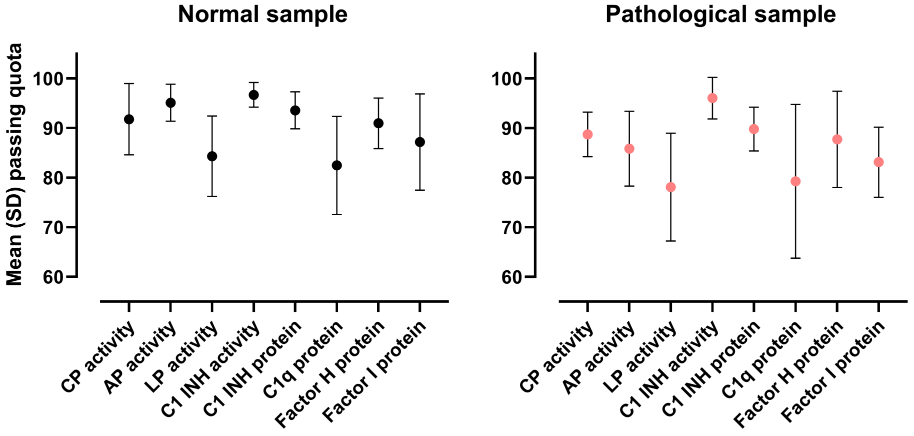


FIGURE 3
Overview of the complement function and proteins (EQA246) EQA results. Data shown are means with standard deviation of the results obtained in the 12 surveys between 2017 and 2022.

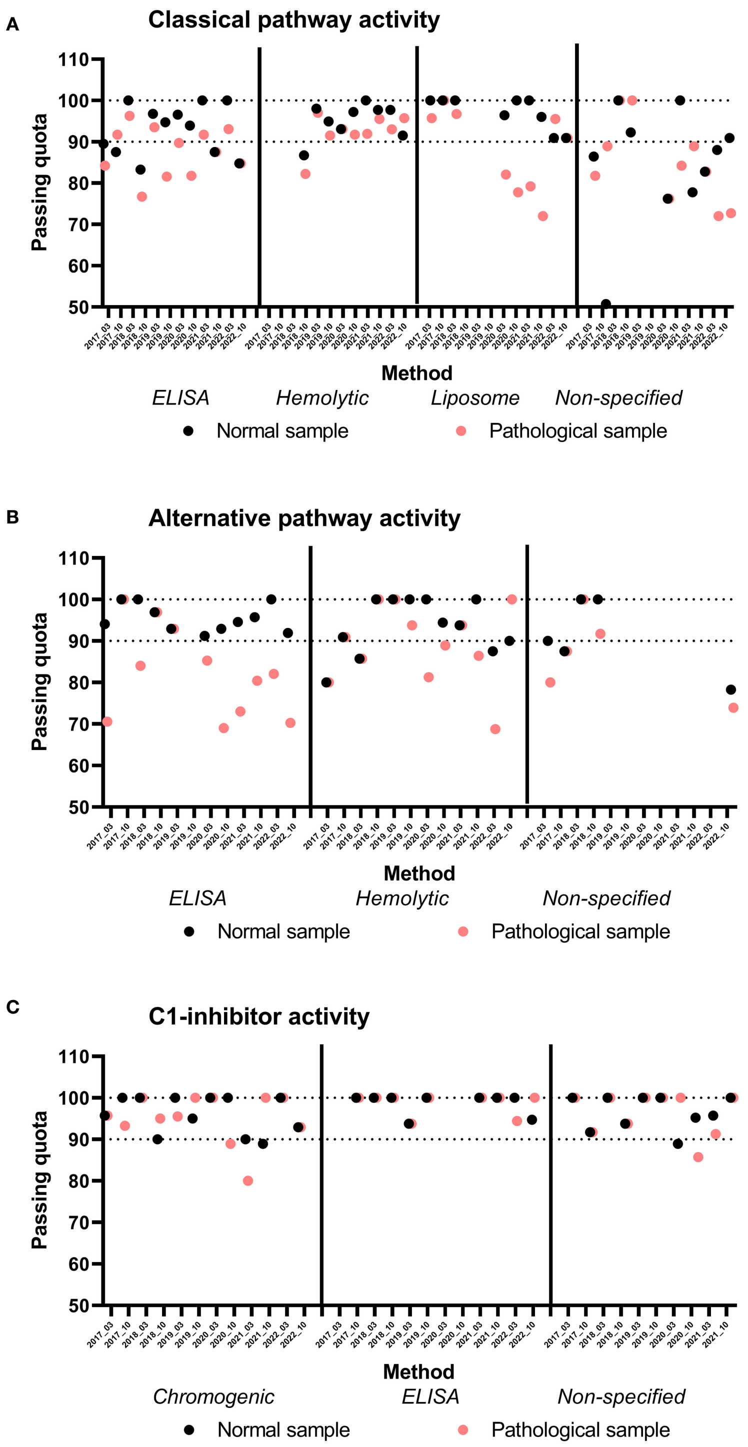


FIGURE 4
Results of classical **(A)**, alternative pathway functional activity **(B)**, and C1-inhibitor activity **(C)**, split by assay methods and EQA surveys. Passing quota of the indicated samples are plotted for indicated EQA surveys and assay groups. Note, that groups with less than 8 participants are not analyzed.

participants with less than four participations showed highly variable results, with passing quota between 0% and 100%.

Finally, as summarized in Table 4, for IgG, IgA and IgM anti-C1-inhibitor autoantibodies (EQA250) interest was generally low (13 laboratories) with only six of thirteen taking part in more than three EQA rounds. None of them performed well for IgG, but six of six succeeded for IgA, and four of six for IgM.

Discussion

The need for a collaborative effort to monitor and improve the quality of complement testing was recognized in 2010. The successful introduction of an EQA was established only six years later as a complex and widely available program for which the results could be entered online. Such a program for analyzing the

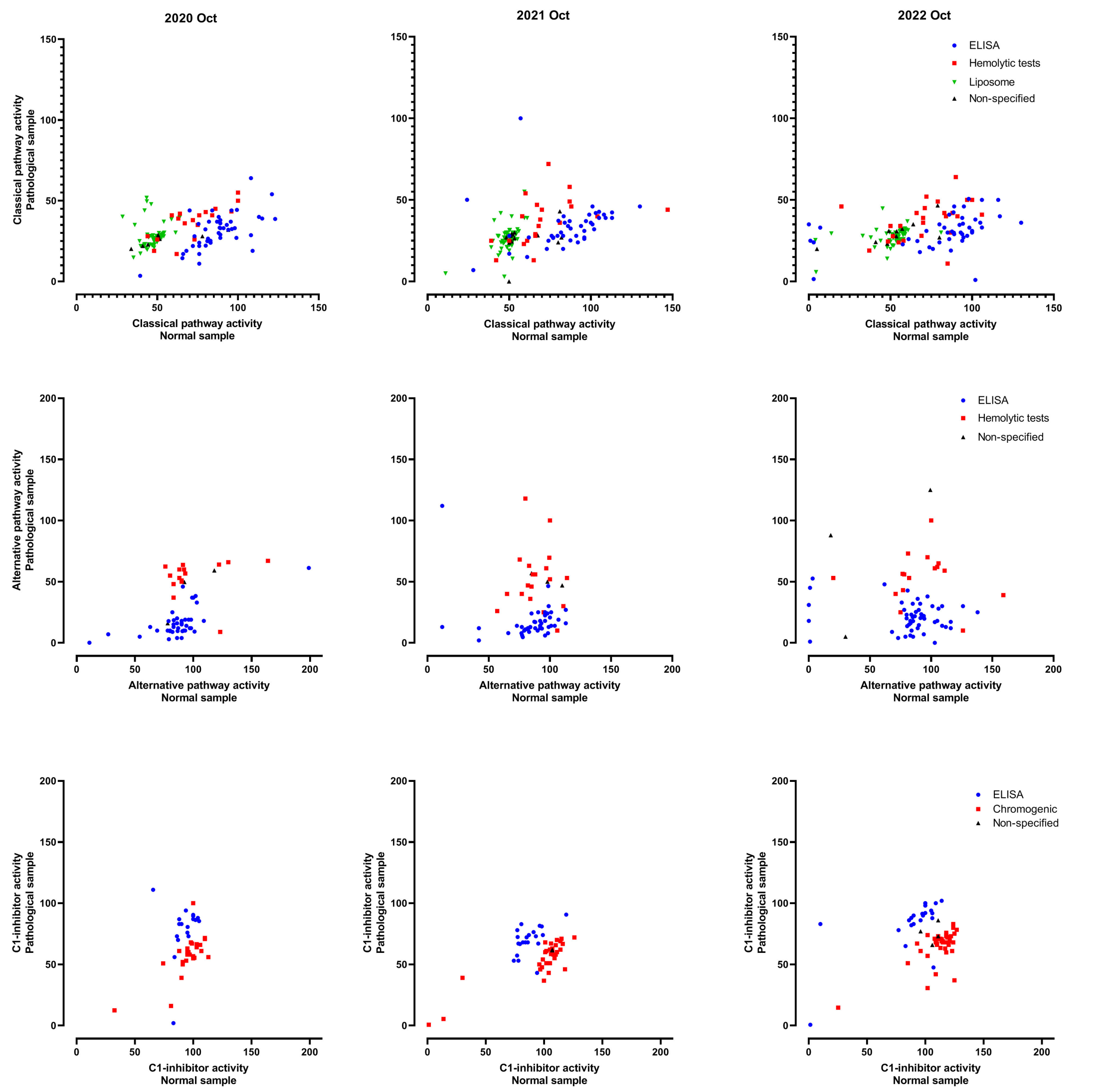


FIGURE 5
Individual results of classical and alternative pathway functional activity, and C1-inhibitor activity. Data shown are activity results (%) of the normal and the pathological sample, assay methods are indicated by the symbols/colors.

highly labile complement system presented a number of challenges, but by joining the expertise of the International Complement Society (ICS) with the knowledge and infrastructure of INSTAND e.V., a successful program could be initiated. This program not only gave an overview on the current state of complement diagnostic testing performance, but also provides the information necessary to improve complement testing procedures.

Our data demonstrates that the passing quota, across the assessments, is higher for normal samples than pathological samples. Looking first at functional analysis, the success rate for pathological samples in CP activity assessment demonstrates variations between the testing methods. Even for a given specific method, the passing quota varied between the years. For the first year the passing rate for both the normal and pathological samples analyzed by the hemolytic assay was below 90%, but in all subsequent years this method was most consistent, particularly for the pathological samples (Figure 4). Results of the ELISA initially had a lower passing quota for the pathological sample, but improved in recent years. This improvement may be attributed to the growing experience with this newer method. On the other side, the lower consistency of the non-specified method could be in part attributable to the lower number of laboratories reporting in this category. For the liposomal assay for CP activity in four rounds of testing (2020 and 2021), the passing quota for the pathogenic sample was ≤80%. These results are consistent with other publications suggesting that this method of measuring CP function is ideal for measuring low level activity (42, 43). However, it should be noted that in more recent rounds the passing quota for the pathological samples improved to greater than 90%. This is important because this method is more commonly used by standard hospital laboratories. Furthermore, a tighter clustering reflects less lab-to-lab variability in the reported results, an important consideration for comparability of testing results between laboratories (Figure 5).

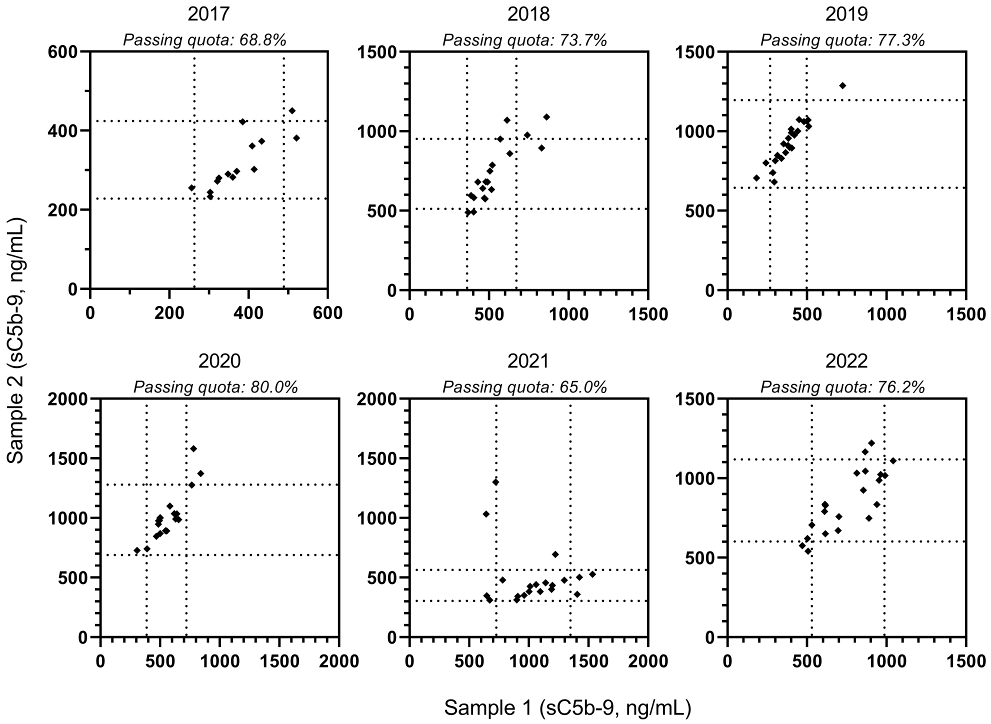


FIGURE 6
Individual results of terminal pathway activation marker (sC5b-9) levels, measured by ELISA assay of Quidel. Data shown are sC5b-9 results of the low and the high concentration samples, as obtained in the past six EQA surveys. Dotted lines indicate acceptance limits for the samples; passing quota (both of the results "passed") of the collective is indicated above the figure for the different years.

AP activity is measured by fewer laboratories with less available testing methods. The passing quota for this analysis was overall lower than for the CP function especially for pathological samples. With the increasing recognition of its importance for disease development and drug monitoring the demand for this testing will certainly increase. Multiple AP specific therapeutic inhibitors are currently in Phase 3 clinical trials (44).

The overall passing quota for C1-INH function testing was higher and more consistent than those for CP or AP function measurements. Demand driven by need to follow therapeutic treatment may be part of the reason for the higher passing quota in testing C1-INHF. There are still method to method differences as is shown in Figure 5, where the reported results do group by method, but the spread of results is much tighter than for CP or AP function. Another contributing factor to the higher passing rate of C1-INHF may be the relative simplicity of testing the function of just one complement regulator, rather than a whole pathway.

The complexity of the complement CP and AP function tests is both their strength diagnostically, but also a potential cause of the observed variability. Their strength comes from the ability to evaluate eleven (CP) or nine (AP) different components in one test, respectively (45). For normal activity, all the components must be present and active. Any therapeutic inhibition along these pathways results in low or abnormal levels, also unraveling the complexities that arise from measuring the function of so many proteases at once. The relationship between protein concentration and activity of an individual component also relates to their drastically different concentrations in serum (from >1 g/L for C3 to 0.1 g/L for C1q, for example). Certain components are rate limiting and due to the stepwise nature of complement activation with several amplification steps the relationship between the component levels of the test serum and its activity is not strictly linear, but rather follows a Von Krogh equation (46). All current methods for measuring complement CP and AP function were developed for testing errors in inborn immunity and not for evaluating therapeutic inhibition of the complement system as now required. As more complement targeting drugs are approved, this may add pressure to the need on complement function testing.

In addition to the use of those functional assays, measurement of activation fragments is also growing in interest in response to the needs related to therapeutic interventions of the complement system. It is for this reason that the soluble membrane attack complex (sMAC, sC5b-9) has the highest participation rate of any of the activation markers. This complex has been proposed as a marker to better reflect that a patient responds to complement inhibition, or to assess if complement activation is causative for the clinical presentation (47–49). However, the utility of measuring sC5b-9 is not undisputed (50), probably also due an inconsistency of the measurements. In Figure 6, sC5b-9 data over six years of EQA assessment are shown. This analyte is only part of the October assessment and only reported by a minority of the participating laboratories. The passing quota of both the normal and pathologic samples reached 80% only once (2020) whereas most years it was

TABLE 1A Participation, passing quota and laboratory performance in the external quality assurance program EQA245 for anti-C1q IgG autoantibody (reagent: Orgentec).

Laboratory	2017 MARCH	2017 OCT	2018 MARCH	2018 OCT	2019 MARCH	2019 OCT	2020 MARCH	2020 OCT	2021 MARCH	2021 OCT	2022 MARCH	2022 OCT	Performance	Participation, total
1	🟦												100%	1
2		🟦	🟦							🟦		🟦	100%	4
3		🟦		🟥									50%	2
4						🟦	🟦	🟦	🟦	🟦	🟦	🟦	**100%**	7
5											🟦	🟦	100%	2
6		🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**91%**	11
7				🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**89%**	9
8	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥		🟦			**89%**	9
9	🟦	🟦	🟦	🟥	🟦	🟦		🟦	🟦	🟦	🟦	🟦	**91%**	11
10					🟦		🟦			🟦		🟦	100%	4
11		🟦		🟦	🟦	🟦							100%	4
12	🟦	🟦	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦			**80%**	10
13		🟥			🟥	🟦	🟦	🟥	🟦	🟦	🟦	🟦	67%	9
14		🟦			🟦		🟦	🟦		🟦	🟦	🟦	**100%**	7
15		🟦				🟦	🟦	🟦	🟦	🟦	🟦	🟦	**100%**	8
16		🟦		🟦	🟦				🟦	🟦	🟦	🟦	**100%**	7
17									🟦	🟦	🟦	🟦	100%	4
18		🟦		🟦	🟦	🟦	🟦	🟥	🟥	🟦	🟦	🟦	**80%**	10
19	🟦												100%	1
20								🟦		🟦		🟦	100%	3
21				🟥		🟦			🟦	🟦			75%	4
22			🟦	🟥	🟦	🟦		🟦	🟦	🟦	🟦	🟦	**89%**	9
23		🟦											100%	1
24		🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**91%**	11
25	🟦	🟦	🟦	🟥	🟦	🟦		🟦	🟦	🟦	🟦		**91%**	10
26						🟦	🟦	🟦	🟦	🟦			100%	5
27			🟦		🟦		🟦		🟦		🟦		100%	5
28						🟦		🟥		🟦		🟦	75%	4
29											🟦	🟦	100%	2
30												🟦	100%	1
Passing quota	100.0%	93.3%	88.9%	30.8%	93.3%	100.0%	100.0%	75.0%	93.8%	100.0%	100.0%	100.0%		
Participation, total	6	15	9	13	15	16	13	16	16	21	16	19		

Participation: Number of submitted results in the period of 2017-2022. Performance: Percentage of submitted results in the target range for both of the samples (dark blue). Any result out of target range (light red), lack of results (white). Passing quota: Performance of laboratories in the indicated surveys. Bold facing: laboratories with at least six participation and at least 80% performance.

TABLE 1B Participation, passing quota and laboratory performance in the external quality assurance program EQA245 for anti-C1q IgG autoantibody (reagent: in-house or INOVA).

Laboratory	2017 MARCH	2017 OCT	2018 MARCH	2018 OCT	2019 MARCH	2019 OCT	2020 MARCH	2020 OCT	2021 MARCH	2021 OCT	2022 MARCH	2022 OCT	Performance	Participation, total
31							🟦						100%	1
32				🟥	🟦		🟦	🟦	🟦				80%	5
33					🟥								0%	1
34							🟦						100%	1
35	🟦												100%	1
36			🟥		🟦	🟦	🟦		🟦	🟦	🟦	🟦	**88%**	8
37		🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**91%**	11
38		🟦											100%	1
39			🟦				🟦				🟦		100%	3
40	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**92%**	12
41							🟦	🟦	🟦			🟦	100%	4
42											🟦	🟦	100%	2
43	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**92%**	12
44										🟥			0%	1
45									🟦		🟦		100%	2
46		🟦					🟦						100%	2
47	🟦			🟥									50%	2
48	🟦												100%	1
49	🟦	🟥	🟦	🟥	🟦		🟦	🟦	🟦	🟦	🟦	🟥	73%	11
50		🟦											100%	1
51		🟦		🟥									50%	2
52	🟦		🟦		🟦		🟦		🟦		🟦		**100%**	6
53	🟦		🟦		🟦		🟦		🟦		🟦		**100%**	6
54		🟦	🟦		🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**100%**	10
55		🟦		🟦		🟦		🟦		🟦		🟦	**100%**	6
56				🟦		🟦							100%	2
57				🟥		🟦							50%	2
58		🟥		🟦									50%	2
59				🟥									0%	1
60							🟦		🟦		🟥	🟦	75%	4
61		🟦		🟦	🟦	🟦		🟦	🟦	🟥	🟦	🟦	**89%**	9
62						🟥		🟥		🟥		🟥	0%	4
63					🟦								100%	1
64					🟦		🟦	🟥	🟦	🟦	🟦	🟦	**86%**	7
65					🟦		🟦	🟥	🟦	🟦	🟦		**83%**	6
66								🟥	🟦				50%	2
67							🟦						100%	1
68									🟦	🟦	🟦	🟦	100%	4
69											🟥	🟦	50%	2
Passing quota	100.0%	83.3%	88.9%	30.8%	92.9%	90.0%	100.0%	69.2%	100.0%	76.9%	88.2%	86.7%		
Participation, total	8	12	9	13	14	10	18	13	17	13	17	15		

Participation: Number of submitted results in the period of 2017-2022. Performance: Percentage of submitted results in the target range for both of the samples (dark blue). Any result out of target range (light red), lack of results (white). Passing quota: Performance of laboratories in the indicated surveys. Bold facing: laboratories with at least six participation and at least 80% performance.

TABLE 2 Participation, passing quota and laboratory performance in the external quality assurance program EQA248 for C3-nefritic factor (C3Nef).

Laboratory	2016	2017	2018	2019	2020	2021	2022	Performance	Participation, total
1				🟥	🟥	🟥	🟦	25%	4
2	🟦	🟦	🟦	🟥	🟥	🟦	🟦	71%	7
3				🟥	🟥	🟥	🟥	0%	4
4		🟥	🟥	🟥	🟥	🟦	🟥	17%	6
5	🟦	🟥	🟥	🟥	🟦	🟦	🟦	57%	7
6	🟦	🟦	🟦	🟥	🟥	🟦	🟦	71%	7
7		🟦		🟦	🟦	🟦	🟦	**100%**	5
8		🟥	🟥					0%	2
9	🟦							100%	1
10	🟦	🟦	🟦	🟦	🟥	🟦	🟦	**86%**	7
11	🟦	🟦	🟦	🟦	🟦	🟥	🟦	**86%**	7
12	🟦	🟦	🟦	🟦	🟥	🟦	🟦	**86%**	7
13	🟥							0%	1
14	🟦	🟦	🟦			🟦		**100%**	4
15	🟦	🟥	🟥	🟦	🟥	🟥	🟦	43%	7
16	🟥		🟥	🟦		🟦		50%	4
17							🟥	0%	1
18		🟥	🟥	🟦		🟥	🟦	40%	5
19							🟦	100%	1
20			🟥	🟥	🟦	🟥	🟥	20%	5
21			🟥	🟥	🟥	🟥	🟥	0%	5
22			🟦		🟦		🟦	100%	3
23			🟥	🟥				0%	2
24						🟥		0%	1
25				🟦	🟥	🟥	🟦	50%	4
26						🟥	🟥	0%	2
27					🟥	🟥	🟦	33%	3
28						🟥		0%	1
Passing quota	81.8%	58.3%	43.8%	47.1%	31.3%	42.9%	70.0%		
Participation, total	11	12	16	17	16	21	20		

Participation: Number of submitted results in the period of 2016-2022. Performance: Percentage of submitted results in the target range for both of the samples (dark blue). Any result out of target range (light red), lack of results (white). Passing quota: Performance of laboratories in the indicated years. Bold facing: laboratories with at least four participations and at least 80% performance.

TABLE 3 Participation, passing quota and laboratory performance in the external quality assurance program EQA249 for anti-Factor H IgG autoantibody.

Laboratory	2016	2017	2018	2019	2020	2021	2022	Performance	Participation, total
1		🟦	🟦	🟦	🟦	🟥	🟦	**83%**	6
2	🟦	🟦	🟦	🟥	🟦	🟦	🟦	**86%**	7
3		🟦	🟦	🟦	🟦	🟦		**100%**	5
4	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**100%**	7
5			🟦					100%	1
6	🟦	🟦			🟦	🟦		**100%**	4
7			🟥					0%	1
8	🟦							100%	1
9	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**100%**	7
10	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**100%**	7
11	🟦	🟦	🟦	🟥	🟦	🟦	🟦	**86%**	7
12	🟦							100%	1
13	🟦	🟦	🟦			🟦		**100%**	4
14	🟦	🟦	🟦	🟦	🟦	🟥	🟦	**86%**	7
15		🟦	🟦					100%	2
16	🟦							100%	1
17	🟥							0%	1
18			🟦		🟦	🟦	🟥	75%	4
19		🟦	🟦	🟦			🟦	**100%**	4
20	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**100%**	7
21		🟦	🟦			🟦		**100%**	3
22		🟦	🟦	🟦	🟦	🟦		**100%**	5
23		🟥		🟦			🟦	67%	3
24		🟦		🟦	🟦			100%	3
25					🟦		🟦	100%	2
26				🟥	🟥	🟦	🟦	50%	4
27			🟦					100%	1
28							🟦	100%	1
29					🟥			0%	1
30					🟦			100%	1
31						🟦	🟦	100%	2
32							🟦	100%	1
Passing quota	92.3%	94.1%	94.4%	78.6%	88.2%	87.5%	93.8%		
Participation, total	13	17	18	14	17	16	16		

Participation: Number of submitted results in the period of 2016-2022. Performance: Percentage of submitted results in the target range for both of the samples (dark blue). Any result out of target range (light red), lack of results (white). Passing quota: Performance of laboratories in the indicated years. Bold facing: laboratories with at least four participations and at least 80% performance.

TABLE 4 Participation, passing quota and laboratory performance in the external quality assurance program EQA250 for anti-C1-inhibitor autoantibodies.

Laboratory	2016	2017	2018	2019	2020	2021	2022	Performance	Participation, total
	IgG								
1	🟦	🟦	🟦	🟦	🟥	🟥	🟥	57%	7
2	🟥	🟦			🟥	🟥	🟥	20%	5
3		🟥	🟦	🟦	🟥			50%	4
4		🟥	🟥					0%	2
5	🟥							0%	1
6	🟦	🟦	🟦	🟦	🟥	🟥	🟥	57%	7
7	🟥	🟥	🟦	🟦	🟦	🟦	🟦	71%	7
8			🟦	🟦	🟦			100%	3
9	🟦							100%	1
10	🟥	🟥	🟦	🟥		🟥	🟥	17%	6
11						🟦	🟦	100%	2
12					🟦			100%	1
13						🟦	🟦	100%	2
Passing quota	42.9%	42.9%	85.7%	83.3%	42.9%	42.9%	42.9%		
Participation, total	7	7	7	6	7	7	7		
	IgA								
1	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**100%**	7
2	🟦	🟦			🟦	🟦	🟦	**100%**	5
3		🟦	🟦	🟦	🟦			**100%**	4
4									0
5	🟦							100%	1
6	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**100%**	7
7	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**100%**	7
8			🟦	🟦	🟦			100%	3
9	🟦							100%	1
10	🟦	🟦	🟦	🟦			🟦	**100%**	5
11									0
12					🟦			100%	1
13						🟦	🟥	50%	2
Passing quota	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	83.3%		

(Continued)

TABLE 4 Continued

Laboratory	2016	2017	2018	2019	2020	2021	2022	Performance	Participation, total
Participation, total	7	6	6	6	7	5	6		
	IgM								
1	🟦	🟦	🟦	🟥	🟦	🟦	🟦	**86%**	7
2	🟦	🟦			🟦	🟥	🟦	**80%**	5
3		🟦	🟦	🟦	🟥			75%	4
4									0
5	🟦							100%	1
6	🟦	🟦	🟦	🟦	🟦	🟦	🟦	**100%**	7
7	🟦	🟦	🟦	🟦	🟦	🟦	🟥	**86%**	7
8	🟥							0%	1
9			🟦	🟦	🟦			100%	3
10	🟦	🟥	🟦	🟦			🟥	60%	5
11									0
12					🟦			100%	1
13						🟦	🟥	50%	2
Passing quota	85.7%	83.3%	100.0%	83.3%	85.7%	80.0%	50.0%		
Participation, total	7	6	6	6	7	5	6		

Participation: Number of submitted results in the period of 2016-2022. Performance: Percentage of submitted results in the target range for both of the samples (dark blue).Any result out of target range (light red), lack of results (white). Passing quota: Performance of laboratories in the indicated years. Bold facing: laboratories with at least four participations and at least 80% performance. Laboratory numbers in the three parts of the table indicate the same participants.

only about 70% although testing was done in just one assay purchased from one manufacturer. The reason for this low passing quota is currently unclear, but may be explained by lab performance, or lot variations of the kit reagents, and low number of participants. It is unfortunate that the calibrator aimed to serve complement activation product assays (51) could not get more interest or acceptance in the past years, and the use is limited to a few laboratories. To this end, laboratories with divergent results are encouraged, as part of the EQA participation, to review their testing if their results do not receive a passing quota.

Of diagnostic importance is the measurement of autoantibodies which is hampered by the limited availability of sufficient quantities of appropriate samples for the complement related autoantibodies. As samples are taken from different patients in different years, variations in EQA results – probably also related to different methods applied- are not surprising. Specifically, the results for anti-C1q IgG autoantibodies in 2018 and 2020 demonstrated a notably lower level of agreement. This was true also for laboratories and methods that were otherwise highly consistent. The specific reasons for this discordance warrants further investigation, especially with reference to clinical presentation. In reviewing these results it is also important to keep in mind the low numbers of participating laboratories for some of these tests. When there are only a few laboratories reporting, individual results may have more impact on the overall passing quota.

The results presented for the complement autoantibodies exemplify an important practical shortage related to this field, i.e. how feasible it is for a small/new laboratory to introduce determination of for example anti-FH, anti-C1INH or C3Nef, as a new parameter. This difficulty is traced back to multiple factors, among which lack of international calibrators and control materials, and lack of commercial interest in these small diagnostic fields are the most important. The quality assessment group/committee already started to produce and share such control materials. One purpose of our article is to attract potential industrial partners and to improve the feasibility of the kit development.

A potential limitation of the current analysis is related to the fact that the test materials offered in this program are not exactly similar to that ones used in the daily routine work. This fact is related largely to logistic and financial aspects, however, during the initial elaboration of the program in the years between 2010 and 2016 efforts were done in the laboratories of the authors to identify the circumstances (in terms of recovery, stability and homogeneity) that are at the same time logistically feasible and technically sound. This is why lyophilization was introduced for three of the programs, and sample shipment at ambient temperature was accepted. However, these efforts make it not unnecessary to perform additional local control in the participating laboratories for preanalytical issues, while testing true routine samples.

Similar to other attempts undertaken to improve diagnostic immunology testing by the International Union of Immunological Societies (IUIS), the efforts of the ICS and INSTAND eV for complement testing is an important step towards improving its quality and standardization. With this view on the current state of testing our data are considered to empower the individual laboratories with knowledge for improvement of their performance, otherwise not available. At this more mature state of the EQA testing these data can facilitate international efforts to investigate how the current methods can be improved for better test results. Without such EQA data, it would be harder to identify the problems that need to be addressed, and any improvement would hardly be measurable.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by Hungarian Ethical Review Agency (ETT-TUKEB). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

MK: Conceptualization, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing, Investigation. AF-A: Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. EB: Methodology, Validation, Writing – original draft, Writing – review & editing, Resources. SG: Data curation, Formal analysis, Methodology, Project administration, Validation, Writing – original draft, Writing – review & editing. NW: Conceptualization, Data curation, Investigation, Methodology, Project administration, Validation, Writing – original draft, Writing – review & editing. ZP: Methodology, Validation, Writing – original draft, Writing – review & editing, Conceptualization, Formal analysis, Project administration, Supervision, Visualization.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We gratefully acknowledge the help of Márta Kókai, Éva Zsuzsanna Szendrei, and Lászlóné Kertész for preparation of complement EQA samples.

Conflict of interest

Authors SG and NW are employees of INSTAND e.V., Düsseldorf, Germany.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1368399/full#supplementary-material

References

1. Kareem S, Jacob A, Mathew J, Quigg RJ, Alexander JJ. Complement: Functions, location and implications. *Immunology*. (2023) 170:180–92. doi: 10.1111/imm.13663

2. Ricklin D, Hajishengallis G, Yang K, Lambris JD. Complement: a key system for immune surveillance and homeostasis. *Nat Immunol*. (2010) 11:785–97. doi: 10.1038/ni.1923

3. Barrington R, Zhang M, Fischer M, Carroll MC. The role of complement in inflammation and adaptive immunity. *Immunol Rev*. (2001) 180:5–15. doi: 10.1034/j.1600-065X.2001.1800101.x

4. Sarma JV, Ward PA. The complement system. *Cell Tissue Res*. (2011) 343:227–35. doi: 10.1007/s00441-010-1034-0

5. Kohl J. The role of complement in danger sensing and transmission. *Immunol Res*. (2006) 34:157–76. doi: 10.1385/IR:34:2

6. Nauser CL, Sacks SH. Local complement synthesis-A process with near and far consequences for ischemia reperfusion injury and transplantation. *Immunol Rev*. (2023) 313:320–6. doi: 10.1111/imr.13144

7. Morgan BP, Gasque P. Extrahepatic complement biosynthesis: where, when and why? *Clin Exp Immunol*. (1997) 107:1–7. doi: 10.1046/j.1365-2249.1997.d01-890.x

8. Yadav MK, Maharana J, Yadav R, Saha S, Sarma P, Soni C, et al. Molecular basis of anaphylatoxin binding, activation, and signaling bias at complement receptors. *Cell*. (2023) 186:4956–4973 e21. doi: 10.1016/j.cell.2023.09.020

9. Xie CB, Jane-Wit D, Pober JS. Complement membrane attack complex: new roles, mechanisms of action, and therapeutic targets. *Am J Pathol*. (2020) 190:1138–50. doi: 10.1016/j.ajpath.2020.02.006

10. Merle NS, Church SE, Fremeaux-Bacchi V, Roumenina LT. Complement system part I - molecular mechanisms of activation and regulation. *Front Immunol*. (2015) 6:262. doi: 10.3389/fimmu.2015.00262

11. Schmidt CQ, Lambris JD, Ricklin D. Protection of host cells by complement regulators. *Immunol Rev*. (2016) 274:152–71. doi: 10.1111/imr.12475

12. Zipfel PF, Skerka C. Complement regulators and inhibitory proteins. *Nat Rev Immunol*. (2009) 9:729–40. doi: 10.1038/nri2620

13. Prohaszka Z, Kirschfink M, Frazer-Abel A. Complement analysis in the era of targeted therapeutics. *Mol Immunol*. (2018) 102:84–8. doi: 10.1016/j.molimm.2018.06.001

14. West EE, Kemper C. Complosome - the intracellular complement system. *Nat Rev Nephrol*. (2023) 19:426–39. doi: 10.1038/s41581-023-00704-1

15. Ekdahl KN, Teramura Y, Hamad OA, Asif S, Duehrkop C, Fromell K, et al. Dangerous liaisons: complement, coagulation, and kallikrein/kinin cross-talk act as a linchpin in the events leading to thromboinflammation. *Immunol Rev*. (2016) 274:245–69. doi: 10.1111/imr.12471

16. Schmidt CQ, Schrezenmeier H, Kavanagh D. Complement and the prothrombotic state. *Blood*. (2022) 139:1954–72. doi: 10.1182/blood.2020007206

17. Ozen A, Comrie WA, Ardy RC, Dominguez Conde C, Dalgic B, Beser OF, et al. CD55 deficiency, early-onset protein-losing enteropathy, and thrombosis. *N Engl J Med*. (2017) 377:52–61. doi: 10.1056/NEJMoa1615887

18. Brodszki N, Frazer-Abel A, Grumach AS, Kirschfink M, Litzman J, Perez E, et al. European society for immunodeficiencies (ESID) and european reference network on rare primary immunodeficiency, autoinflammatory and autoimmune diseases (ERN RITA) complement guideline: deficiencies, diagnosis, and management. *J Clin Immunol*. (2020) 40:576–91. doi: 10.1007/s10875-020-00754-1

19. Grumach AS, Kirschfink M. Are complement deficiencies really rare? Overview on prevalence, clinical importance and modern diagnostic approach. *Mol Immunol*. (2014) 61:110–7. doi: 10.1016/j.molimm.2014.06.030

20. Ricklin D, Lambris JD. Complement in immune and inflammatory disorders: pathophysiological mechanisms. *J Immunol*. (2013) 190:3831–8. doi: 10.4049/jimmunol.1203487

21. Skattum L. Clinical complement analysis-an overview. *Transfus Med Rev*. (2019) 33:207–16. doi: 10.1016/j.tmrv.2019.09.001

22. Ricklin D, Mastellos DC, Reis ES, Lambris JD. The renaissance of complement therapeutics. *Nat Rev Nephrol*. (2018) 14:26–47. doi: 10.1038/nrneph.2017.156

23. Mollnes TE, Jokiranta TS, Truedsson L, Nilsson B, Rodriguez de Cordoba S, Kirschfink M. Complement analysis in the 21st century. *Mol Immunol*. (2007) 44:3838–49. doi: 10.1016/j.molimm.2007.06.150

24. Ling M, Murali M. Analysis of the complement system in the clinical immunology laboratory. *Clin Lab Med*. (2019) 39:579–90. doi: 10.1016/j.cll.2019.07.006

25. Frazer-Abel A, Sepiashvili L, Mbughuni MM, Willrich MA. Overview of laboratory testing and clinical presentations of complement deficiencies and dysregulation. *Adv Clin Chem*. (2016) 77:1–75. doi: 10.1016/bs.acc.2016.06.001

26. Ekdahl KN, Persson B, Mohlin C, Sandholm K, Skattum L, Nilsson B. Interpretation of serological complement biomarkers in disease. *Front Immunol*. (2018) 9:2237. doi: 10.3389/fimmu.2018.02237

27. Prohaszka Z, Frazer-Abel A. Complement multiplex testing: Concept, promises and pitfalls. *Mol Immunol*. (2021) 140:120–6. doi: 10.1016/j.molimm.2021.10.006

28. Matola AT, Jozsi M, Uzonyi B. Overview on the role of complement-specific autoantibodies in diseases. *Mol Immunol*. (2022) 151:52–60. doi: 10.1016/j.molimm.2022.08.011

29. Prohaszka Z, Nilsson B, Frazer-Abel A, Kirschfink M. Complement analysis 2016: Clinical indications, laboratory diagnostics and quality control. *Immunobiology*. (2016) 221:1247–58. doi: 10.1016/j.imbio.2016.06.008

30. Brandwijk R, Michels M, van Rossum M, de Nooijer AH, Nilsson PH, de Bruin WCC, et al. Pitfalls in complement analysis: A systematic literature review of assessing complement activation. *Front Immunol*. (2022) 13:1007102. doi: 10.3389/fimmu.2022.1007102

31. Mollnes TE, Garred P, Bergseth G. Effect of time, temperature and anticoagulants on *in vitro* complement activation: consequences for collection and preservation of samples to be examined for complement activation. *Clin Exp Immunol*. (1988) 73:484–8.

32. Yang S, McGookey M, Wang Y, Cataland SR, Wu HM. Effect of blood sampling, processing, and storage on the measurement of complement activation biomarkers. *Am J Clin Pathol*. (2015) 143:558–65. doi: 10.1309/AJCPXPD7ZQXNTIAL

33. van der Pol P, de Vries DK, van Gijlswijk DJ, van Anken GE, Schlagwein N, Daha MR, et al. Pitfalls in urinary complement measurements. *Transpl Immunol*. (2012) 27:55–8. doi: 10.1016/j.trim.2012.06.001

34. Sikkeland LIB, Ueland T, Lund MB, Durheim MT, Mollnes TE. A role for the terminal C5-C9 complement pathway in idiopathic pulmonary fibrosis. *Front Med (Lausanne)*. (2023) 10:1236495. doi: 10.3389/fmed.2023.1236495

35. Zelek WM, Fathalla D, Morgan A, Touchard S, Loveless S, Tallantyre E, et al. Cerebrospinal fluid complement system biomarkers in demyelinating disease. *Mult Scler*. (2020) 26:1929–37. doi: 10.1177/1352458519887905

36. Struglics A, Okroj M, Sward P, Frobell R, Saxne T, Lohmander LS, et al. The complement system is activated in synovial fluid from subjects with knee injury and from patients with osteoarthritis. *Arthritis Res Ther*. (2016) 18:223. doi: 10.1186/s13075-016-1123-x

37. Schick T, Steinhauer M, Aslanidis A, Altay L, Karlstetter M, Langmann T, et al. Local complement activation in aqueous humor in patients with age-related macular degeneration. *Eye (Lond)*. (2017) 31:810–3. doi: 10.1038/eye.2016.328

38. Sonntag J, Brandenburg U, Polzehl D, Strauss E, Vogel M, Dudenhausen JW, et al. Complement system in healthy term newborns: reference values in umbilical cord blood. *Pediatr Dev Pathol*. (1998) 1:131–5. doi: 10.1007/s100249900016

39. Johnson U, Truedsson L, Gustavii B. Complement components in 100 newborns and their mothers determined by electroimmunoassay. *Acta Pathol Microbiol Immunol Scand C*. (1983) 91:147–50.

40. Roach B, Kim Y, Jerome E, Michael AF. Influence of age and sex on serum complement components in children. *Am J Dis Child*. (1981) 135:918–20. doi: 10.1001/archpedi.1981.02130340030011

41. *ISO 13528:2022: Statistical methods for use in proficiency testing by interlaboratory comparison*. Geneva, Switzerland: International Organization for Standardization, ISO (2022). Available at: https://www.iso.org/obp/ui/en/#iso:std:iso:13528:ed-3:v1:en.

42. Willrich MAV, Ladwig PM, Martinez MA, Sridharan MR, Go RS, Murray DL, et al. Monitoring Ravulizumab effect on complement assays. *J Immunol Methods*. (2021) 490:112944. doi: 10.1016/j.jim.2020.112944

43. Gatault P, Brachet G, Ternant D, Degenne D, Recipon G, Barbet C, et al. Therapeutic drug monitoring of eculizumab: Rationale for an individualized dosing schedule. *MAbs*. (2015) 7:1205–11. doi: 10.1080/19420862.2015.1086049

44. West EE, Woodruff T, Fremeaux-Bacchi V, Kemper C. Complement in human disease: approved and up-and-coming therapeutics. *Lancet*. (2023) 403(10424):392–405. doi: 10.1016/S0140-6736(23)01524-6

45. Walport MJ. Complement. First of two parts. *N Engl J Med*. (2001) 344:1058–66. doi: 10.1056/NEJM200104053441406

46. Rosse WF, Dacie JV. Immune lysis of normal human and paroxysmal nocturnal hemoglobinuria (PNH) red blood cells. I. The sensitivity of PNH red cells to lysis by complement and specific antibody. *J Clin Invest*. (1966) 45:736–48. doi: 10.1172/JCI105388

47. Schoettler ML, Carreras E, Cho B, Dandoy CE, Ho VT, Jodele S, et al. Harmonizing definitions for diagnostic criteria and prognostic assessment of transplantation-associated thrombotic microangiopathy: A report on behalf of the european society for blood and marrow transplantation, american society for transplantation and cellular therapy, asia-pacific blood and marrow transplantation group, and center for international blood and marrow transplant research. *Transplant Cell Ther*. (2023) 29:151–63. doi: 10.1016/j.jtct.2022.11.015

48. Chauvet S, Hauer JJ, Petitprez F, Rabant M, Martins PV, Baudouin V, et al. Results from a nationwide retrospective cohort measure the impact of C3 and soluble

C5b-9 levels on kidney outcomes in C3 glomerulopathy. *Kidney Int*. (2022) 102:904–16. doi: 10.1016/j.kint.2022.05.027

49. Qi J, Wang J, Chen J, Su J, Tang Y, Wu X, et al. Plasma levels of complement activation fragments C3b and sC5b-9 significantly increased in patients with thrombotic microangiopathy after allogeneic stem cell transplantation. *Ann Hematol*. (2017) 96:1849–55. doi: 10.1007/s00277-017-3092-9

50. Wehling C, Amon O, Bommer M, Hoppe B, Kentouche K, Schalk G, et al. Monitoring of complement activation biomarkers and eculizumab in complement-mediated renal disorders. *Clin Exp Immunol*. (2017) 187:304–15. doi: 10.1111/cei.12890

51. Bergseth G, Ludviksen JK, Kirschfink M, Giclas PC, Nilsson B, Mollnes TE. An international serum standard for application in assays to detect human complement activation products. *Mol Immunol*. (2013) 56:232–9. doi: 10.1016/j.molimm.2013.05.221

TYPE Original Research
PUBLISHED 17 May 2024
DOI 10.3389/fmolb.2024.1395410

OPEN ACCESS

EDITED BY
Matteo Becatti,
University of Firenze, Italy

REVIEWED BY
Anna Frances Lau,
National Institutes of Health (NIH), United States
Olga Perovic,
National Institute of Communicable Diseases (NICD), South Africa

*CORRESPONDENCE
Stefan Ziesing,
✉ ziesing.stefan@mh-hannover.de

RECEIVED 03 March 2024
ACCEPTED 29 April 2024
PUBLISHED 17 May 2024

CITATION
Lindenberg M, Waldmann S, Suerbaum S, Schlüter D and Ziesing S (2024), External quality assessment schemes in bacteriology support public health in Germany—results from 2006 to 2023.
Front. Mol. Biosci. 11:1395410.
doi: 10.3389/fmolb.2024.1395410

External quality assessment schemes in bacteriology support public health in Germany—results from 2006 to 2023

Marc Lindenberg[1], Sabine Waldmann[1], Sebastian Suerbaum[2,3,4], Dirk Schlüter[1,5,6] and Stefan Ziesing[1,6]*

[1]Institute for Medical Microbiology and Hospital Epidemiology, Hannover Medical School, Hannover, Germany, [2]German Center for Infection Research (DZIF), Munich, Germany, [3]Max von Pettenkofer Institute, Faculty of Medicine, Ludwig-Maximilians-Universität, Munich, Germany, [4]National Reference Center for Helicobacter Pylori, Munich, Germany, [5]German Center for Infection Research (DZIF), Partner Site Hannover-Brunswick, Hannover, Germany, [6]Management of External Quality Assessment Schemes Bacteriology, Instand e.V., Düsseldorf, Germany

External Quality Assessment schemes (EQAS) are mandatory to ensure quality standards in diagnostic methods and achieve laboratory accreditation. As host institution for two German culture-based bacteriology EQAS (RV-A and RV-B), we investigated the obtained data of 590 up to 720 surveys per year in RV-A and 2,151 up to 2,929 in RV-B from 2006 to 2023. As educational instruments, they function to review applied methodology and are valuable to check for systemic- or method-dependent failures in microbiology diagnostics or guidelines. Especially, containment of multi-resistant bacteria in times of rising antibiotic resistance is one major point to assure public health. The correct identification and reporting of these strains is therefore of high importance to achieve this goal. Moreover, correct antimicrobial susceptibility testing (AST) *per se* is important for selecting appropriate therapy, to restrict broad-spectrum antibiotics and minimize resistance development. The reports of participating laboratories displayed a high level of correct identification results in both schemes with mostly consistent failure rates around 2.2% (RV-A) and 3.9% (RV-B) on average. In contrast, results in AST revealed increasing failure rates upon modification of AST requirements concerning adherence to standards and subsequent bacterial species-specific evaluation. Stratification on these periods revealed in RV-A a moderate increase from 1.3% to 4.5%, while in RV-B failure rates reached 14% coming from 4.3% on average. Although not mandatory, subsequent AST evaluation and consistent reporting are areas of improvement to benefit public health.

KEYWORDS

microbiology, bacteriology, external quality assessment, public health, antimicrobial susceptibility testing (AST), identification methods bacteriology

1 Introduction

Conventional culture-based identification of bacteria and subsequent antimicrobial susceptibility testing (AST) remain the gold standard and represent the largest part of bacteriological diagnostics in medical microbiology, although molecular biological methods have and will further improve the identification of bacterial pathogens. However, at present, AST as a central task of every diagnostic microbiological-bacteriological laboratory can only be performed adequately by culture-based techniques but not by molecular biological methods including whole genome sequencing (Jorgensen and Ferraro, 2009; Turnidge et al., 2023). Due to the outstanding importance for the detection of infections and selection of suitable therapeutic options based on AST - especially in times of constantly increasing antibiotic resistance–the applied methods are subject to not only internal laboratory quality assessment but also external quality assessment, which is mandatory in Germany. The public health system is also dependent on assured and constantly evolving quality in bacteriology especially concerning i) reliable and fast identification for reporting of notifiable pathogens, ii) rapid and reproducible AST in accordance with standards to inform clinicians about safe and efficient treatment options and to prevent unnecessary usage of broad-spectrum substances, iii) and an up-to-date and uniform nomenclature, as well as antibiotic-susceptibility assessment standards such as EUCAST and CLSI to assure correct communication between key players of the healthcare system.

Different national laws and guidelines oblige microbiology laboratories to participate in External Quality Assessment schemes (EQAS). In Germany, the Federal Medical Council issues these binding guidelines for all medical laboratories (RiliBÄK) (Bundesaerztekammer, 2023). Reference institutions including Instand e.V. manage these EQAS in collaboration with host diagnostic microbiological laboratories. For laboratories performing bacteriology diagnostics, successful participation in EQAS, at least once a year, is a prerequisite to receive reimbursement of costs for diagnostic procedures with the respective cost bearers. In Germany, INSTAND e.V. has been performing EQAS in bacteriology with fast-growing organisms since 2006 with the Institute for Medical Microbiology and Hospital Epidemiology of Hannover Medical School, Hannover, Germany as host institution. The host institution acts as scientific management partner with selection of suitable bacterial strains, production of specimens, evaluation, and commenting of results for each survey. Instand e.V. organizes the surveys with respect to the shipment of specimens, both nationally and internationally, recording the results and providing them to the host laboratory for final evaluation. Successful participation in EQAS is a prerequisite to obtaining accreditation, as stated in the International Standard ISO 15189:2022 (ISO, 2022).

In Germany, diagnostic bacteriology is performed by specialized laboratories but also by outpatient practitioners who provide diagnostics for their specialty, and here urologists are by far the largest group by numbers. The German guidelines consider the diagnostic differences leading to two different EQA schemes. Bacteriology "Ringversuch A" (RV-A), directed to specialized microbiology laboratories and sent out twice a year with five bacterial samples, and "Ringversuch B" (RV-B), directed to outpatient practitioners and sent out four times a year with three probes containing urogenital pathogens or commensals but not restricted to bacteria. In both schemes, slow-growing bacteria like mycobacteria, which are subject to separate EQAS, are excluded. Besides direct quality assurance, other beneficial aspects of EQA are that the host institution issues a certificate upon successful participation, which is mandatory to hand costs to the respective cost-bearers. In addition, the EQA host institution is obliged to report abnormalities to the respective authorities for instance to the Federal Institute for Drugs and Medical Devices. Of note, every EQAS round is also a test for the issued diagnostic guidelines for instance the breakpoint tables. Finally, EQAS are educational and can spread new knowledge on nomenclature, current epidemiology, clinical relevance of microorganisms, newly described resistance mechanisms, and phenotypic appearances that may lead to misinterpretation. The appended commentaries in the result reports are highly valuable in spreading knowledge. Within this study, we analyzed a highly consistent 288 to 364 laboratory reports per survey in RV-A and a more varying 392 to 940 per survey in RV-B as a large and representative database (Figure 1A).

2 Materials and Methods

2.1 Identification part in EQAS bacteriology

In RV-A directed to specialized microbiology laboratories, the host laboratory sends out five specimens of bacterial strains twice a year with one specimen per year as a mixture of two strains. Participants need to identify strains on genus and species level and obtain one point for correct identification per level. To pass this category at least 80% of all points need to be gained per survey. The 80% cut-off value has already been defined since at least 1992 and has not been changed with the takeover of the EQAS by the current host laboratory.

In RV-B, performed four times a year and directed to laboratories performing bacteriology within the scope of their respective profession, which is overwhelmingly urology, the focus is on urogenital pathogens and commensals. Here, three specimens are sent but are not strictly limited to bacteria but can also contain yeast strains, without consequent susceptibility testing. Four of the six points must be gained in this category to pass in RV-B per survey.

Reference results are obtained from a consortium of 16 highly qualified microbiology laboratories, which are referred to as target value laboratories (TVL) from here on. Not all TVL take part in every survey. TVL are suggested by the host laboratory and have to be accepted by the Federal Medical Council. Most of them have acted as TVL for more than 2 decades.

2.2 Antimicrobial susceptibility testing (AST) part in EQAS bacteriology

From a table of 16 antibiotics for RV-A and 15 antibiotics for RV-B, the participants have to choose and report those suitable to treat the identified bacterial species in accordance with the used AST standard. For each specimen a minimum count of antibiotics (approximately three-quarters of the maximum number assessable) to be tested is defined by the host institution of the EQAS, dependent on the number of antibiotics evaluable with respect to the utilized AST standard. Participants need to test the identified bacteria and report interpreted results as susceptible S), intermediate respectively susceptible at increased dosing (I, the latter definition valid for EUCAST since 2019), or resistant R). For every substance, a full point is gained by meeting one interpretation of the

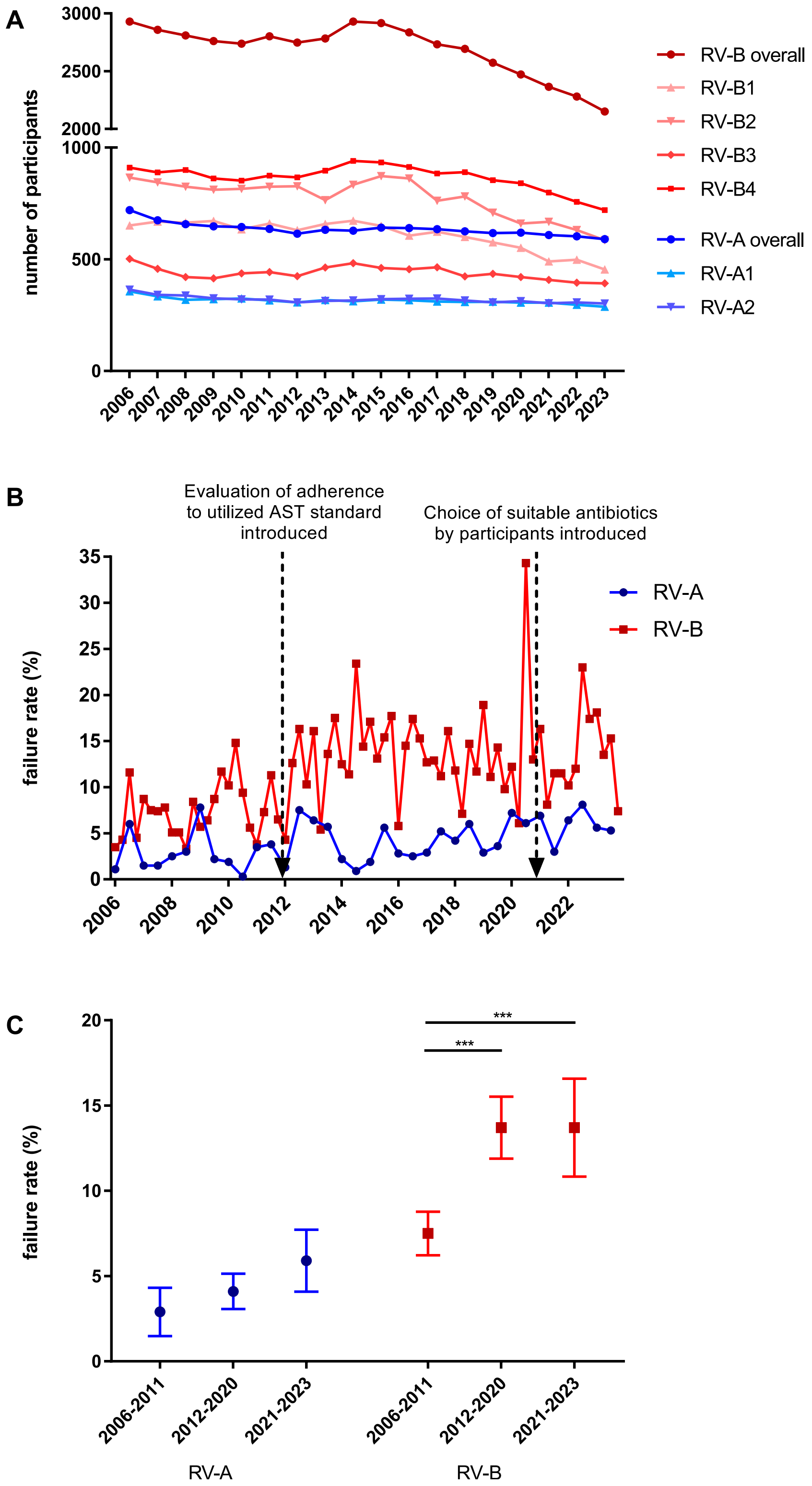


FIGURE 1
Analysis of participants and passing rates in bacteriology EQAS RV-A and RV-B. **(A)** Number of participants in both EQA schemes from 2006 to 2023 for the respective annual dates. **(B)** Failure rates in RV-A (blue) and RV-B (red) EQAS at the respective dates with red arrows indicating time points of the described modifications in the EQAS. **(C)** Overall failure rates in RV-A and RV-B categorized for the periods between the aforementioned modifications in the EQAS. Depicted are mean ± SD for 12 (2006–2011), 18 (2012–2020), and 6 (2021–2023) data points in RV-A and 24 (2006–2011), 36 (2012–2020), and 12 (2021–2023) in RV-B.

set point range, while half a point is granted in case of "minor errors"; I instead of R, for example, (Turnidge et al., 2023). In RV-A and RV-B, both the given minimum number of antibiotics and 85% of all points for correct results are needed to pass. AST is performed in parallel to the participants by the TVL three times to account for technical variability, which can be method-dependent. TVL report only one final result to the host laboratory and are asked to deliver results for at least two combinations of technique–disk diffusion or MIC determination–and AST standard. Based on these values the set point range is determined. Depending on the scattering of TVL results the target value is set usually to one level as S, I, or R. Following a defined algorithm, in case of broader scattering more than one level might be accepted.

2.3 Timeline of modifications

Initially in 2006, when the Institute of Medical Microbiology and Hospital Epidemiology of Hannover Medical School took over the management of the EQAS, participants were required to identify the strain on genus and species level and test susceptibility against at least six out of eight defined antibiotics by disk diffusion according to the German Institute for Standardization (Deutsches Institut für Normung, DIN) standard. At first, we re-defined the required AST panel with respect to the strain identified as being Gram-positive or Gram-negative. From 2012 onwards, participants were required to report the utilized AST standard, while the ones of DIN, Clinical and Laboratory Standards Institute (CLSI), and European Committee on Antimicrobial Susceptibility Testing (EUCAST, 2024) were accepted. While DIN was excluded in 2014 as being outdated and discontinued, EUCAST modified by recommendations of the national antibiotic susceptibility testing committee Germany (NAK) for certain substances was accepted from 2016 onwards (EUCAST and EUCAST + NAK were summarized for data analysis in this paper). Reported results in AST were evaluated in correlation to the reported and utilized standard.

2.4 Data collection

Participants are asked to report their results on standardized questionnaires in each EQAS round.

While up to 2019 paper-based reports had to be handed in by the participants, from then onwards an online form is mandatory. This online form made it feasible to obtain additional data on pathogenicity, extended bacterial typing, reporting obligations, detected mechanism of resistance, and, as a German specialty, multi-resistant phenotypes in Gram-negative rods which are used for management in hospital hygiene and, in part, have to be reported to the public health authorities.

2.5 Data analysis

We analyzed the reported results of participating laboratories from 2006 until 2023. As the EQA definitions and requirements changed over time, different analysis topics span different time frames defined by changes in the requirements for susceptibility testing. From 2006 to 2011: Disk diffusion according to DIN only. From 2012 to 2021: Reporting the utilized guideline for result interpretation (DIN, CLSI, EUCAST, EUCAST plus NAK), both disk diffusion and MIC techniques possible. From 2021 onwards: Participants have to select the antibiotic substances to be reported in accordance with the utilized AST standard.

2.6 Statistics

GraphPad Prism Version seven was used to determine significance of results. Figure legends describe statistical tests run on respective data sets. One-way-Analysis of variance (ANOVA)-test was used if not indicated differently and means are given as ± *s*.d. with p values considered significant as follows: * = $p < 0.05$; ** = $p < 0.005$ and *** = $p < 0.0005$.

3 Results

3.1 Analysis of participants and passing rates in bacteriology EQA schemes RV-A and RV-B

For public health, a high standard of microbiological diagnostics and a solid data basis is required and EQAS evaluate this for the participating laboratories. Since 2006, we recognized an almost constant number of participants in RV-A, directed at specialized microbiological laboratories at both time points per year. In contrast, the number of participants in RV-B varied greatly on the four annual dates. The last date of each year had by far the highest number of participants. From 2006 to 2023, the overall failure rates in these EQAS were in the range of 0.3%–8.1% in RV-A (mean 4.0% ± 2.23%) and 3.3%–34.3% in RV-B (mean 11.6% ± 5.39%). Failure rates increased for periods following modifications to the EQAS with respect to AST evaluation (details in Material and Methods) (Figure 1B). Further analyzing the results of the different periods with increasing demands on participants, we found them significantly increased for RV-B and trending in the same direction on a lower overall level for RV-A (Figure 1C). Additionally, failure rates were substantially higher in RV-B as compared to RV-A illustrating better diagnostic quality of RV-A participants according to the EQAS criteria.

3.2 Identification of bacteria, development of identification success, and methods used

To determine the reasons for increased failure rates upon EQAS modification, we further analyzed failure rates in identification and AST separately as both had to be passed by participants. Correct identification of bacteria is a prerequisite for correct AST considering species-specific breakpoints. The analysis of identification results from 2009 to 2023 showed no significant changes over time for both RV-A and RV-B and also between RV-A and RV-B with failure rates as low as 2.1% (±1.6%) in RV-A and 3.9% (±3.30%) in RV-B in this EQAS category (Figure 2A). However, as successful identification rates varied between different

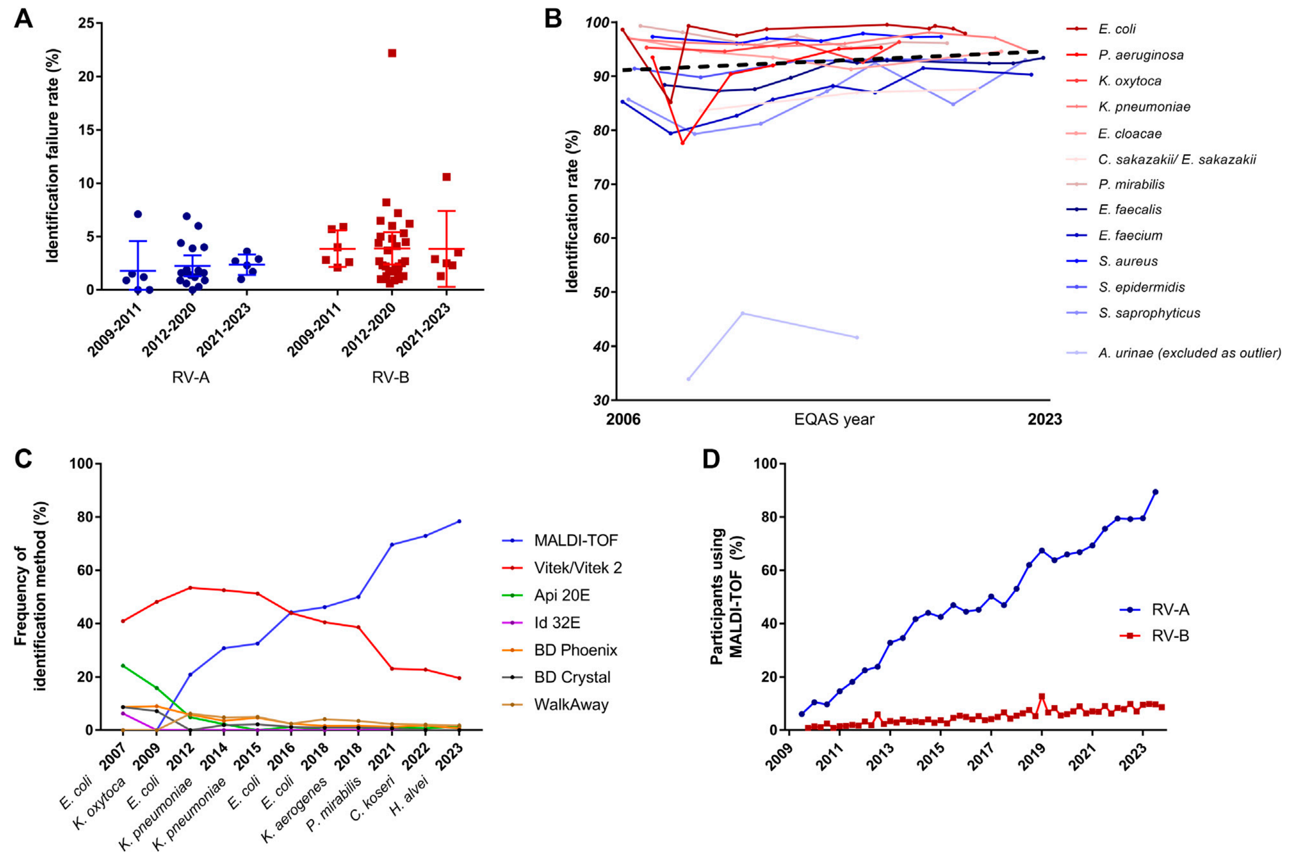


FIGURE 2
Analysis of identification rates over time. **(A)** Failure rates in identification in RV-A and RV-B categorized in periods with regard to modifications in the EQAS. Dots represent respective overall failure rates per EQAS survey. **(B)** Identification rate for bacterial species being sent out at least three times from 2009 until 2023 in RV-B. Gram-negatives depicted in shades of red, gram-positives in shades of blue according to the legend in the graph, while the dotted line represents the regression line over all data points excluding rates for *A. urinae*. Regression analysis gave a slope of 0.049 (95% CI: 0.009–0.108). **(C)** Analysis of utilized identification methods for Enterobacterales in RV-A for the indicated time points. **(D)** Frequency of MALDI-TOF as identification method among RV-A (blue) and RV-B (red) participants.

bacterial species, we checked for improvements on a bacterial species-dependent level. In RV-A considerable improvements in the accuracy of identification of rarely detected and challenging bacterial species were already shown by the EQAS of the time since their introduction in 1982 (Schaal, 1994). Since 2006, we found significant improvements only for a comparatively small number of species (Table 1). For some species, occasionally a slight decline in identification rates was found compared to previous EQAS rounds. In most of these cases, affected strains were part of a germ mixture consisting of two bacteria, where the potential problem of retrieval may add to the increased failure rate.

Moreover, we analyzed the identification success for bacterial species being sent out repeatedly (in both RV-A and RV-B. While participants in RV-A were overall more successful than in RV-B, for common urogenital pathogens an overall high standard of identification rates was observed (Table 2). However, bacterial species rarely causing urinary tract infections but need to be identified in accordance with diagnostic guidelines were more challenging for participants in RV-B (Table 3).

As the failure rates in identifications among participants in RV-A were very low, we focused on RV-B results to track changes over time. To account for the educational aspect of the EQAS, we analyzed the identification rate of bacterial species in RV-B being sent out three or more times between 2006 and 2023. We observed a slight trend to increased accuracy in identification rates, however, the slope of the fitted regression line was not significantly different from zero (p = 0.09) (Figure 2B).

Nonetheless, we asked for methodological improvements over time. Therefore, we analyzed developments in the participant's identification methods used for Enterobacterales identification during different EQAS rounds. While RV-A laboratories overwhelmingly changed to rely on Matrix-associated-laser-desorption-ionization and time of flight (MALDI-TOF) analysis, a technique considered to be fast and of high accuracy (Dingle and Butler-Wu, 2013), this technique is still not widely used by RV-B participants (Figures 2C, D).

3.3 Adherence to nomenclature

In view of consistent reporting of diagnostic microbiological results to clinicians and health authorities, correct identification and also the use of current terminology is a desirable goal. Therefore, we studied

TABLE 1 Identification rate for selected species in RV- A. Displayed species are selected due to their relevance with respect to guideline adherence, taxonomic changes, frequency of isolation, and culture conditions. (*: strain has been part of a germ mixture).

Species	Date	Rate [%]	Date	Rate [%]	Date	Rate [%]
Arcobacter butzleri	1–2009	60.6	1–2019	73.5		
Bacillus pumilus	1–2011	77.5	1–2022	84.8		
Bacteroides fragilis	1–2006	95.2	1–2021	94.4	2–2023	94
Campylobacter jejuni	1–2007	91.6	2–2020	92.7		
Clostridioides (Clostridium) difficile	2–2006	99.2	2–2015	94.7	1–2019	94.8
Clostridium tertium	2–2006	88.4	1–2022	91.6		
Corynebacterium belfantii, C. diphteriae complex, C. rouxii; toxin-negative	2–2014	97.8	1–2023	96.5		
Cronobacter (Enterobacter) sakazakii	2–2007	100	1–2021	99.7		
Cutibacterium (Propionibacterium) acnes	2–2008	95.3	1–2009	87.1*	2–2018	92.7
Eikenella corrodens	2–2007	97.3	1–2023	97.9		
Finegoldia magna	1–2011	95.6	1–2021	92.5*		
Granulicatella (Abiotrophia) adiacens	2–2012	89.2	1–2020	91.5		
Haemophilus influenzae	2–2011	98.4	1–2016	98.1		
Kingella kingae	2–2008	93.2	2–2022	93.5		
Klebsiella (Enterobacter) aerogenes (Klebsiella mobilis)	2–2010	98.4	2–2018	99.7		
Listeria monocytogenes	2–2010	99.4	2–2016	98.5	2–2020	98.1
Mammaliicoccus (Staphylococcus) sciuri	1–2019	98.4	1–2023	100		
Micrococcus luteus	2–2007	99.7	1–2015	90.3*		
Pasteurella multocida	1–2012	94.1	1–2013	93.3		
Rahnella aquatilis	1–2008	98.1	1–2017	96.8		
Serratia marcescens	1–2007	99.4	2–2009	96.3*	1–2017	99.4
Staphylococcus (Peptococcus) saccharolyticus	2–2010	94.7	1–2014	67*		
Staphylococcus caprae	1–2017	96.8	1–2022	98.3		
Staphylococcus lugdunensis	1–2008	98.1	1–2016	98.4		
Staphylococcus schleiferi	2–2011	100	2–2018	98.4		
Streptococcus canis	2–2017	81.5	2–2021	89.4		
Streptococcus dysgalactiae	1–2016	93.1	2–2019	93.2		
- dysgalactiae equisimilis		68.8		64.2		
Streptococcus gallolyticus sp. *Gallolyticus*	2–2009	98.2	1–2014	99.4		
Vibrio vulnificus	2–2008	98.5	1–2019	97.1		
Weeksella virosa (CDC group IIf, Flavobact. sp?)	2–2012	88.9	1–2020	81.4		
Yersinia enterocolitica	2–2008	99.7	1–2014	99.7	1–2022	98.6

results from EQAS strains with changes in taxonomy (Table 4). For participants, there are usually no drawbacks when adhering to outdated names, as both new and old names are accepted in the EQAS, and hence this topic is not recapitulated in the failure rate analysis. Categorizing the time since the publication of the new name and the respective EQAS round, we found a significant correlation between the updated taxonomy being reported and a period greater than 5 years since the renaming (Figure 3).

3.4 Results of antimicrobial susceptibility testing (AST)

As the overall increase in failure rates was not attributable to the identification rates, we analyzed the AST results of participants. The specialist microbiological laboratories in RV-A documented a high level of quality in AST, while RV-B participants showed poorer accuracy rates, which also scattered over a wider range. The two

TABLE 2 Comparison of identification rates in RV-A and RV-B for frequent urogenital species.

Species	RV-A			RV-B		
	Surveys [n]	Mean [%]	Range [%]	Surveys [n]	Mean [%]	Range [%]
Gram-negative						
Escherichia coli	6	99.7	99.3–100	10	97.4	85.2–99.3
Pseudomonas aeruginosa	4	98.5	96.4–99.7	10	92.3	77.6–96.9
Klebsiella pneumoniae	2	99.1	98.4–99.7	8	96.3	94.5–98.1
Proteus mirabilis	4	99.8	99.1–100	7	96.9	95.2–99.3
Enterobacter cloacae	2	97.6	96.1–99.1	5	94.2	91.3–97.1
Gram-positive						
Enterococcus faecalis	3	98.9	98.4–99.4	10	91.0	87.3–93.4
Enterococcus faecium	1	98.7	-	8	86.3	79.4–91.5
Staphylococcus aureus	6	99.6	99.2–100	7	97.0	96.0–97.9
Staphylococcus saprophyticus	1	98.7	-	7	86.2	79.3–93.0
Staphylococcus epidermidis	1	98.8	-	5	92.0	89.8–93.1

TABLE 3 Identification rates for rarely found species in RV-B.

Species	Date	Rate [%]	Date	Rate [%]	Date	Rate [%]	Date	Rate [%]	Date	Rate [%]
K. oxytoca	1–2007	95.3	2–2010	94.6	2–2013	96.2	1–2016	92.6	3–2017	96.3
Klebsiella aerogenes	3–2016	-	2–2019	37.9						
"Enterobacter" aerogenes		92.5		57.5						
Cronobacter sakazakii	2–2009	1.2	4–2015	56.7	4–2020	76.1				
"Enterobacter" sakazakii		82.4		30.3		15.0				
Pantoea agglomerans	3–2009	82.9	3–2012	86.6	4–2022	91.1				
Serratia marcescens	2–2008	95.0	4–2012	94.0	3–2021	91.7				
Corynebakterium urealyticum	2–2008	44.7	3–2012	31.6	3–2017	37.9	2–2023	58.6		
Staphylococcus lugdunensis	2–2016	48.8	1–2022	52.2						
Micrococcus luteus	2–2010	83.6								
Aerococcus urinae	4–2008	33.9	1–2011	46.1	4–2015	41.6				
Actinobaculum (Actinotignum) schaalii	3–2020	16.2								
Lactobacillus	1–2007	41.5	1–2012	55.4						
rhamnosus		5.5		8.3						

considerable changes to the requirements in this part of the EQAS led to increased failure rates in both series, with considerably stronger effects in RV-B (Figure 4A).

With respect to the requirements of the public health system, the performance of the participants in the identification of resistance mechanisms is of particular interest; especially as the initiation of hygiene measures to prevent pathogen spread depends on these results. While not obligatory and not evaluated in the EQAS - as no international standard is applicable - participants had the opportunity to indicate identified bacteria and phenotypic AST combinations with the respective acronym. The reliability of detecting and reporting oxacillin resistance in *Staphylococcus aureus* (MRSA), vancomycin resistance in enterococci (VRE), or Extended-Spectrum-Betalactamase (ESBL) expression in Enterobacterales was high in RV-A (data not shown). However, in RV-B, only a minority of strains were reported with the respective acronyms, even though phenotypically characterized as resistant. Some showed an increase in reporting over time anyway (Table 5).

In Germany, carbapenem resistance due to carbapenemases, especially in Enterobacterales, has only been an epidemiological

TABLE 4 Reported names for bacteria with changes in taxonomy in RV-A.

RV-A	Taxonomy		
	Former name	New name	year of renaming
2–2007: *Aggregatibacter (Haemophilus) aphrophilus*	75.7	17.3	2006
2–2007: *Pantoea (Enterobacter) agglomerans*	1.5	94.4	1989
1–2009: *Cupriavidus (Ralstonia, Wautersia) pauculus*	17.4	73.3	2004
2–2010: *Staphylococcus (Peptococcus) saccharolyticus*	0	94.7	1981
2–2010: *Enterobacter aerogenes (Klebsiella mobilis)*	0	98.4	1960
2–2011: *Raoultella (Klebsiella) planticola*	0.6	81.3	2001
2–2012: *Granulicatella (Abiotrophia) adiacens*	0.3	88.9	2000
2–2012: *Moraxella (Branhamella) catarrhalis*	0	99.0	1968
2–2016: *Actinobaculum (Actinotignum) schaalii*	13.0	73.1	2015
2–2017: *Paeniclostridium (Clostridium) sordellii*	88.0	1.2	2016
2–2018: *Cutibacterium (Propionibacterium) acnes*	63.6	29.1	2016
2–2018: *Klebsiella (Enterobacter) aerogenes*	57.0	42.7	2017
1–2020: *Weeksella virosa* (CDC group IIf, *Flavobacterium*)	0.3	81.7	1986
1–2020: *Granulicatella (Abiotrophia) adiacens*	0.3	91.8	2000
1–2021: *Cronobacter (Enterobacter) sakazakii*	1.3	98.7	2008
1–2021: *Finegoldia magna (Peptostreptococcus)*	2.0	92.5	1999
1–2021: *Pantoea (Enterobacter) agglomerans*	1.6	98.0	1989
2–2021: *Raoultella (Klebsiella) ornithinolytica*	1.0	97.7	2001
2–2022: *Delftia (acidovorans) tsuruhatensis*	92.5	2.3	2003
1–2023: *Empedobacter (Wautersiella) falsenii*	1.4	59.4	2014
2–2023: *Lacticaseibacillus (Lactobacillus) rhamnosus*	41.7	55.3	2020

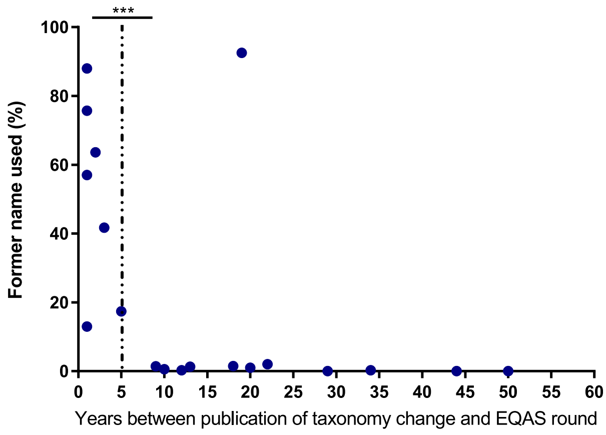


FIGURE 3
Adherence to new taxonomy A Data display the frequency of bacterial species undergoing taxonomy changes being reported with the former name correlated to the time between publication of the new name and the respective EQAS round. The dotted line indicates categorization in more than 5 years since publication and equal or less for the respective data points. Significance was tested by a chi-square test. *** = $p < 0.005$.

problem since 2010 (Albiger et al., 2015). We had a look at the performance of the RV participants over time concerning carbapenemase detection. Table 6 summarizes the results on various carbapenem-resistant bacteria in RV-A and RV-B with green color indicating increases in reporting for repeatedly sent-out strains.

Since the detection of carbapenemases can be challenging, we analyzed the results of a New Delhi carbapenemase (NDM)-producing *Proteus mirabilis* sent out in RV-A1 in 2016. Table 7 compares the documented MIC values reported by target value laboratories (TVL) and categorizes results from participants stratified according to CLSI and EUCAST (Table 7). The MIC values varied widely and both the TVL and the participants rated a similarly high proportion of the tests as "S" or "I", while NDM-carrying bacteria are usually considered phenotypically carbapenem-resistant. We found that more than 30% of EUCAST participants reported the strain meropenem susceptible or intermediate, while this was true for about 10% of CLSI users only. This example illustrates the difficulty in detecting carbapenemase expression solely upon AST and puts the results of only 16.4% of the participants characterizing an NDM- or metallo-ß-lactamase, 19.6% reporting a carbapenemase

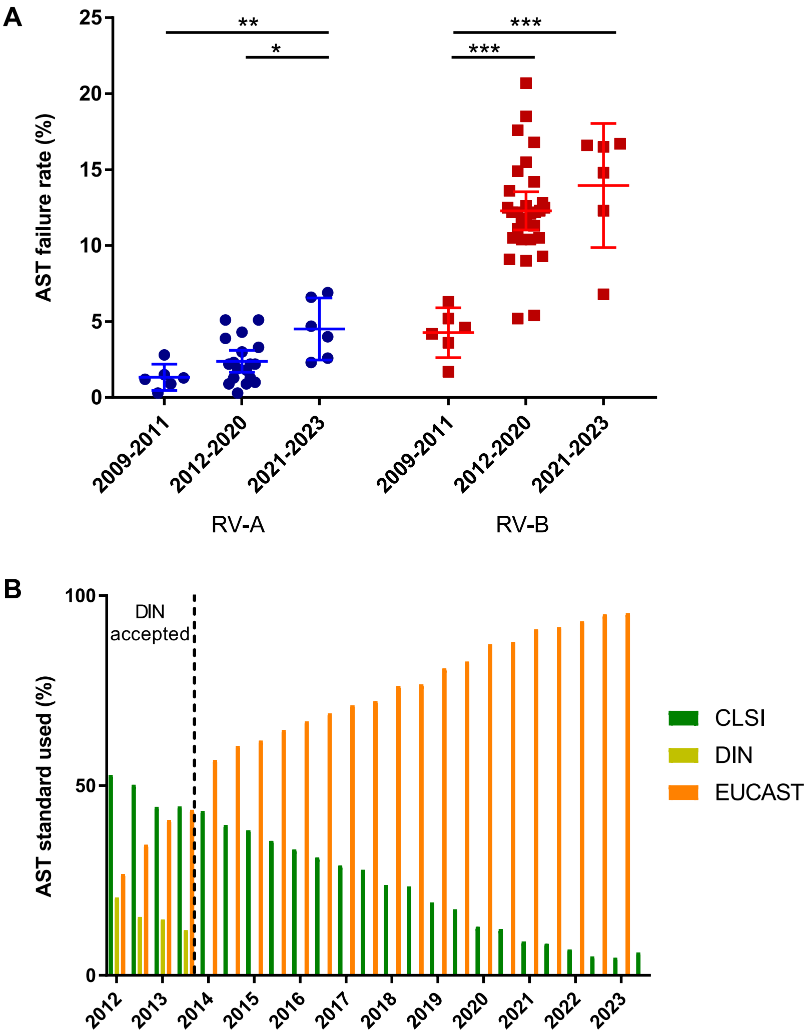


FIGURE 4
Analysis of antimicrobial susceptibility testing (AST). **(A)** Failure rates in AST for RV-A and RV-B categorized in periods with regard to modifications in the EQAS. Depicted are mean ± SD for 6 (2009–2011), 18 (2012–2020), and 6 (2021–2023) data points in RV-A and 6 (2009–2011), 36 (2012–2020), and 6 (2021–2023) in RV-B. **(B)** Frequency of AST standard used by participants in RV-A from 2012 to 2023. Color schemes of the different standards according to the figure legend.

without further characterization and 64.1% of participants indicating no resistance mechanism at all, in perspective. Moreover, evaluation is highly dependent on the utilized AST standard in this case.

To the authors' knowledge, there are no publicly available statistics on the AST standards utilized in laboratories. At least for RV-A, we determined the EUCAST standard to be overwhelmingly applied nowadays, while only a few participants still utilize the CLSI standard being the most applied standard back in 2012 (Figure 4B).

3.5 Adherence to AST recommendations of a German guideline on uncomplicated urinary tract infections

Finally, we investigated how the efforts of medical guidelines were supported by the AST of participants in RV-B. A German urology S3 guideline on uncomplicated urinary tract infections updated in 2017 recommends the antibiotics fosfomycin (single oral dose), nitrofurantoin, nitroxoline, and pivmecillinam for premenopausal women to counteract the constant development of resistance, particularly to fluoroquinolones in the field of urology (Kranz et al., 2018). According to EUCAST, these substances are only to be evaluated in full for *E. coli*. The reported AST for this species in RV-B is therefore a measure of the guideline adherence of the participating laboratories. Reported results for 4 *E. coli* strains sent out in RV-B in 2021 and 2022 were analyzed. We set the most frequently tested antibiotic per round–ciprofloxacin - equal to 100% and found nitrofurantoin (mean value 94.8%), fosfomycin (76.8%), mecillinam (50.6%), and nitroxolin (48.0%) less frequently reported compared to other oral or parenteral antibiotics (Figure 5). Hence, roughly half of RV-B participants did not report on two first-line antibiotics, while the medical society representing most of them recommends their therapeutic use.

TABLE 5 Voluntary reporting on resistance mechanisms for repeatedly sent bacteria in RV-B.

Species: Resistance mechanism	EQAS round	Rate (%)	EQAS round	Rate (%)	EQAS round	Rate (%)	EQAS round	Rate (%)	EQAS round	Rate (%)	EQAS round	Rate (%)
E. coli: ESBL	1–2008	11.7	4–2010	19.3	1–2012	26.3	1–2017	19.7	2–2020	28.9		
K. pneumoniae: ESBL	2–2006	8.4	3–2008	18.3	1–2011	16.5	3–2012	31.8	2–2015	24.8		
P. mirabilis: ESBL			3–2008	16.2	2–2013	18.8						
E. faecium: VRE	1–2008	8.3	4–2010	14.3	2–2012	14.4	4–2014	13.2	3–2016	15.4	1–2023	19.2
S.aureus: MRSA	2–2007	20.2	4–2010	23.7	1–2012	37.8	2–2014	30.0	1–2016	36.1		
S.aureus: MRSA - specified as mecC									1–2018	30.7 1.8	2–2019	5.2 0.8

TABLE 6 Comparison of results in carbapenemase characterization between RV-A and RV-B.

Species: Resistance mechanism	Target value	RV-A			RV-B		
			Reported as. . .			Reported as. . .	
		EQAS round	Any carbapenemase (%)	Target value (%)	EQAS round	Any carbapenemase (%)	Target value (%)
Acinetobacter pittii: GIM-1	MBL	A2-2017	23.8	5.5	N/A		
Citrobacter freundii: VIM-1	MBL	A1-2019	39.0	26.4	B3-2019	11.6	7.1
Escherichia coli: OXA-181	OXA-Type	A2-2016	43.3	17.9	B2-2023	12.5	9.9
Escherichia coli: VIM-1	MBL	A2-2012	54.1	40.7	B1-2013	4.6	2.3
		A2-2020	74.8	59.4	B1-2019	5.9	3.0
					B1-2022	10.2	6.6
Klebsiella pneumoniae: KPC	KPC	N/A			B3-2013	7.3	3.0
					B1-2023	12.8	9.5
Klebsiella pneumonia: NDM	MBL	A1-2015	25.9	16.9	N/A		
Klebsiella pneumonia: OXA-48	OXA-Type	A1-2014	75.3	21.8	B1-2020	13.6	6.4
Klebsiella (Enterobacter) aerogenes: AmpC + porine loss	AmpC + porine loss	A2-2018	8.2	19.3	B2-2019	3.4	2.1
Proteus mirabilis: NDM	MBL	A1-2016	19.6	16.4	B2-2021	10.6	6.7
Pseudomonas aeruginosa: VIM-2	MBL	A2-2016	25.7	17.9	B2-2017	5.0	2.2
					B2-2022	9.0	6.5

3.6 Accession of EQAS management comments

The comments written by the EQAS management for each test round are available for download on the Instand e.V. website after the EQA certificates have been issued. We checked the accession rate of these comments that had been available for at least 3 months at the end of 2023. An average of 19.8% (range 2.9%–56.9%) of participants overall accessed the comments, while the difference between an average of 35.7% of RV-A participants, but only 13.7% of RV-B participants was striking. On the other hand, the comments are probably also of interest to laboratories that did not participate in the respective EQAS: 13.4% of downloads for RV-A and 27.8% for RV-B were made by users who did not participate in the respective EQAS round.

4 Discussion

Analyzing the data obtained by managing the two bacteriology EQAS RV-A and RV-B we recognized the overall high standards in bacteriological diagnostics in Germany with passing rates for

TABLE 7 NDM-expressing *Proteus mirabilis* RV-A1-2016: Variance in meropenem MICs and classification determined by TVL and EQAS participants.

Meropenem							Rate (%)		
MIC (mg/L)	0.5	1	2	4	8	≥16	S	I	R
CLSI-TVL n)	1		1	3	2	9	6.3	6.3	87.5
- participants n)	3		5	70			3.8	6.4	89.7
EUCAST-TVL n)	1		1	3	2	9	12.5	31.3	56.3
- participants n)	7			34		92	5.3	25.6	69.2

specialized laboratories in recent years of 95% or higher (Figure 1B). Previous data from the respective Swiss EQAS from 1992 until 1996 showed comparable results (Siegrist et al., 1998), while analysis of EQA in other countries, especially in developing ones, tended to show lower passing rates. Moreover, three of these studies show educational effects in terms of improvements over time (Chaitram et al., 2003; Perovic et al., 2019; Wattal et al., 2019), while our data, and a study on bacteriology in the Eastern Mediterranean Region, with higher failure rates, hardly detected improvements (Squires et al., 2022). Even if the overall diagnostic accuracy is on a high level, EQAS always have the chance to send out certain bacteria unveiling limitations and directions of improvement; especially by including species, that have come into medical focus only recently and for which correct identification or AST might be not well established in laboratories.

The public health system greatly benefits from this highly reliable diagnostic level, as it serves, as a data basis for epidemiological developments, is crucial in detecting and containing local outbreaks, and guides towards an effective but specific antibiotic therapy. However, the comparability of EQAS data and studies involving human clinical microbiology data, in general, is low and in need of standardized reporting (Turner et al., 2019).

Technological progress is one strengthening aspect in this regard. The identification of bacteria using Matrix Assisted Laser Desorption Ionisation - Time of Flight (MALDI-TOF) has become of great importance for microbiological diagnostics during the study period (Figure 2D). It is considerably faster than biochemical reactions and largely independent of the correct selection of a system suitable for a defined germ spectrum. In addition, the reliability of identification, especially in routine operations, is raised to a previously unknown level. However, the use of the MALDI-TOF method requires a relevant investment that can only be made by larger laboratories. As a consequence, there has been a steady increase in its use in the specialized laboratories participating in RV-A, while its use in the predominantly smaller, specialty-specific, and outpatient-providing laboratories participating in RV-B has remained at a comparatively low level. Only the use of the highly automated Vitek system, which is widely available in the majority of laboratories due to the sensitivity tests carried out with this system, still accounted for a higher proportion of identifications for a relatively long time. Irrespective of the method, our data shows that identification of frequently found bacteria species is more successful compared to rarely found ones

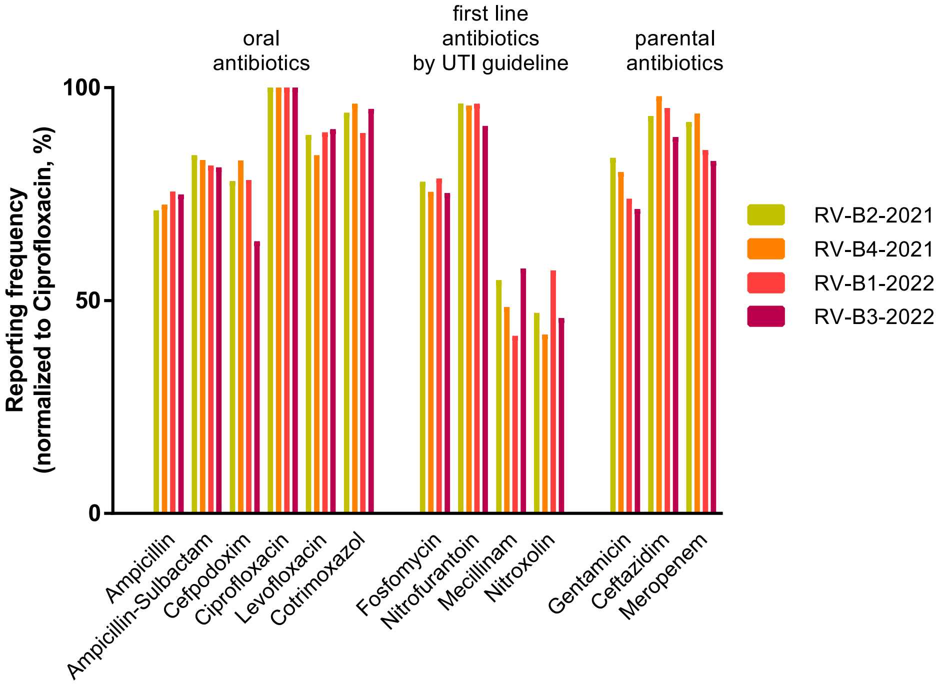


FIGURE 5
Reporting of urology guideline-recommended antibiotics in RV-B. Reporting frequencies of specific antibiotic substances for *Escherichia coli* normalized for ciprofloxacin rates as 100% in four individual RV-B rounds in 2021 and 2022. Substances are grouped into oral and parental antibiotics as well as antibiotics recommended by a German S3 guideline for uncomplicated urinary tract infections as first-line therapy.

recapitulating findings of EQAS in other countries (Wonglumsom et al., 2008; Wattal et al., 2019) (Tables 2, 3). In terms of consistent communication of microbiological reports, the use of current terminology is a desirable goal (Figure 3). A laboratory's constant effort to keep up with this development, which is considerably more dynamic due to molecular biological analyses, is essential for this. In addition, the implementation of the current nomenclature depends to a large extent on the implementation of changes in commercial identification systems by the respective manufacturers.

While the identification of bacterial species remained on a very low failure rate during the observed period, two major modifications in the requirements for successful AST increased the failure rates significantly (Figure 4A). Evidence-based medicine is generating more and more reliable and species-specific data sets on bacterial infections and minimal inhibitory concentration (MIC)-distributions (Leclercq et al., 2013; Kahlmeter, 2015). Correlating these *in vivo* findings with the *in vitro* AST, led to and will further result in an increased amount of breakpoint tables and recommendations in the different standards issued by expert committees. Laboratories have to navigate this development and adhere to a certain standard to evaluate obtained MICs on a good data basis. In this regard, major hurdles leading to failing in the AST category have been recognized. i) Incorrect results for an individual substance caused by technical errors, reagents of insufficient quality, or errors in reading are found to be major aspects not only in this study but also in others (Perovic et al., 2019). ii) Utilizing an outdated version of the reported AST standard leads to failures in evaluations (e.g., categorical result "I", although not defined according to the standard or evaluation of substances no longer seen as applicable for the bacterial species) (Wattal et al., 2019). In this regard, suppliers from commercially available AST systems need to implement updates promptly to enable consistent interpretation not only in terms of AST standards but also concerning updated treatment guidelines (Figure 5). iii) Moreover, substances, for which no specific breakpoints are listed but evaluation can be derived from indicator substances, are not reported (e.g., cefoxitin screen for staphylococci for oxacillin, cefuroxime, ampicillin-sulbactam) and or an incorrect selection of antibiotics is chosen. In particular, since the introduction of antibiotic selection by participants, up to one-third of participants failed the susceptibility testing part due to an insufficient number of antibiotics tested. These findings are in line with the observation of Perovic et al. analyzing the African EQAS concerning AST (Perovic et al., 2019).

In Germany, microbiological laboratories are legally obliged to participate in EQAS. However, they are not obliged to adhere to a specific AST standard. Therefore, the EQAS is the only instance to monitor and evaluate AST utilization (Figure 4B). Nearly all laboratories participating in RV-A applied to the EUCAST standard, which is favorable for public health as this increases the comparability of results and AST evaluations. However, the CLSI standard is still used by a minority of EQAS participants, which can lead to different reports regarding AST. This trend in standards utilization is also found in other European countries (Altorf-van der Kuil et al., 2017; EUCAST international uptake, 2024).

The example of an NDM-expressing *P. mirabilis* indicates the associated variability besides an already existing technical variation of measurements (Table 7). This circumstance makes it difficult to compare the results obtained, at least for individual antibiotics, and thus to communicate them for therapeutic or epidemiological purposes and to projects recording antibiotic resistance developments, like Antibiotic Resistance Surveillance (ARS) managed by Robert Koch Institute (RKI, Berlin, Germany) for instance (ARS - Antibiotika-Resistenz-Surveillance, 2007; Walter et al., 2017). As a consequence, existing AST standards concerning the secure and sensitive detection of epidemiologically relevant resistance mechanisms should be followed thoroughly.

AST standards define how to detect specific resistance mechanisms, however, consistent reporting standards on these are not defined. Hence, reporting of resistance mechanisms can only be assessed in the EQAS on a voluntary basis. For public health applicable standards and subsequent evaluation of the adherence to them in EQA seems as a future goal in light of increasing antimicrobial resistance rates worldwide. Methicillin-resistant staphylococci, vancomycin- or linezolid-resistant enterococci, and ESBL- or carbapenemase-producing Enterobacterales must be identified with the highest degree of certainty to prevent further spread in the healthcare system. While a high degree of reliability is achieved in the determination of the phenotype (evaluation as "R") for prominent resistances in staphylococci (methicillin) and enterococci (vancomycin), the labeling of pathogens as MRSA or VRE is only rarely carried out in RV-B and is likely reflecting daily laboratory reports in an outpatient setting (Table 5).

The aforementioned aspects all indicate the importance of an EQAS focusing on bacteriology for reliable individual diagnostics but also subsequent public health on a bigger scale. Nonetheless, we recognized a decline in the number of participants in RV-B starting around 2014, while RV-A showed more constant participation (Figure 1A). This could be related to the increased quality assurance requirements for laboratories stipulated in the legally binding directive given by the RiliBÄK (Bundesaerztekammer, 2023), but also to an improved offer from specialized laboratories to private practitioners and finally to changes in remuneration leading to a consolidation in the bacterial diagnostics market. The studied EQAS cannot claim to be exhaustive as it is not the only bacteriological EQAS on the market, but the represented number of laboratories both in RV-A and RV-B is representative of the German field of bacteriology.

The issued comments on each EQAS round give the chance to emphasize developments or methodological pitfalls and have thereby the chance to support the spreading of knowledge and indirectly support public health. This fact is recapitulated by the accession not only by participating laboratories in the respective EQAs round but also by registered non-participants.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Individual EQAS results are confidential.

Requests to access these datasets should be directed to SZ, stefan.ziesing@mh-hannover.de.

Author contributions

ML: Formal Analysis, Investigation, Visualization, Writing–original draft, Writing–review and editing. SW: Data curation, Investigation, Project administration, Writing–review and editing. SS: Funding acquisition, Project administration, Supervision, Writing–review and editing. DS: Funding acquisition, Project administration, Supervision, Writing–review and editing. SZ: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Project administration, Supervision, Visualization, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. As the host institution of EQAS Bacteriology, the Institute for Medical Microbiology and Hospital Epidemiology receives reimbursements from Instand e.V. This study was conducted completely independently regarding design, data analysis, and interpretation of results.

Acknowledgments

We thank Sabine Goseberg for excellent administration and data management for the EQAS Bacteriology by Instand e.V. and Gisela Rademacher for technical and administrative expertise in the first 14 years of EQAS Bacteriology managed by our Institute.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Albiger, B., Glasner, C., Struelens, M. J., Grundmann, H., Monnet, D. L., Koraqi, A., et al. (2015). Carbapenemase-producing enterobacteriaceae in europe: assessment by national experts from 38 countries, may 2015. *Eurosurveillance* 20. doi:10.2807/1560-7917.ES.2015.20.45.30062

Altorf-van der Kuil, W., Schoffelen, A. F., de Greeff, S. C., Thijsen, S. F. T., Alblas, H. J., Notermans, D. W., et al. (2017). National laboratory-based surveillance system for antimicrobial resistance: a successful tool to support the control of antimicrobial resistance in The Netherlands. *Eurosurveillance* 22, 17. doi:10.2807/1560-7917.ES.2017.22.46.17-00062

ARS (2007). Antibiotika-resistenz-Surveillance. Available at: https://ars.rki.de/ (Accessed March 1, 2024).

Bundesaerztekammer (2023). Richtlinie der Bundesärztekammer zur Qualitätssicherung laboratoriumsmedizinischer Untersuchungen. Available at: https://www.bundesaerztekammer.de/fileadmin/user_upload/BAEK/Themen/Qualitaetssicherung/_Bek_BAEK_RiLi_BAEK_ONLINE_FINAL_VERS_26_05_2023.pdf (Accessed March 1, 2024).

Chaitram, J. M., Jevitt, L. A., Lary, S., Tenover, F. C., and WHO Antimicrobial Resistancce Group (2003). The world health organization's external quality assurance system proficiency testing program has improved the accuracy of antimicrobial susceptibility testing and reporting among participating laboratories using NCCLS methods. *J. Clin. Microbiol.* 41, 2372–2377. doi:10.1128/JCM.41.6.2372-2377.2003

Dingle, T. C., and Butler-Wu, S. M. (2013). Maldi-tof mass spectrometry for microorganism identification. *Clin. Lab. Med.* 33, 589–609. doi:10.1016/J.CLL.2013.03.001

EUCAST (2024). International uptake. Available at: https://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Statistics/EUCAST_Maps_January_2024.pdf (Accessed March 1, 2024).

ISO 2022. 15189:2022 - medical laboratories — requirements for quality and competence. Available at: https://www.iso.org/standard/76677.html (Accessed February 29, 2024).

Jorgensen, J. H., and Ferraro, M. J. (2009). Antimicrobial susceptibility testing: a review of general principles and contemporary practices. *Clin. Infect. Dis.* 49, 1749–1755. doi:10.1086/647952

Kahlmeter, G. (2015). The 2014 Garrod Lecture: EUCAST - are we heading towards international agreement? *J. Antimicrob. Chemother.* 70, 2427–2439. doi:10.1093/jac/dkv145

Kranz, J., Schmidt, S., Lebert, C., Schneidewind, L., Mandraka, F., Kunze, M., et al. (2018). The 2017 update of the German clinical guideline on epidemiology, diagnostics, therapy, prevention, and management of uncomplicated urinary tract infections in adult patients. Part II: therapy and prevention. *Urol. Int.* 100, 271–278. doi:10.1159/000487645

Leclercq, R., Cantón, R., Brown, D. F. J., Giske, C. G., Heisig, P., Macgowan, A. P., et al. (2013). EUCAST expert rules in antimicrobial susceptibility testing. *Clin. Microbiol. Infect.* 19, 141–160. doi:10.1111/J.1469-0691.2011.03703.X

Perovic, O., Yahaya, A. A., Viljoen, C., Ndihokubwayo, J. B., Smith, M., Coulibaly, S. O., et al. (2019). External quality assessment of bacterial identification and antimicrobial susceptibility testing in african national public health laboratories, 2011–2016. *Trop. Med. Infect. Dis.* 4, 144. doi:10.3390/TROPICALMED4040144

Schaal, K. P. (1994). Modern aspects of microbiological diagnostics A report about the symposium on november 4th, 1993, in Berlin. *LaboratoriumsMedizin* 18, 475–487. doi:10.1515/LABM.1994.18.10.475/PDF

Siegrist, H. H., Pünter-Streit, V., and Von Graevenitz, A. (1998). The Swiss external quality assessment scheme in bacteriology and mycology 1992-1996. *Accredit. Qual. Assur.* 3, 203–207. doi:10.1007/s007690050223

Squires, R. C., Al Abri, A., Al-Rashdi, A., Al Jaaidi, A., Al Harthi, O., Cognat, S., et al. (2022). External quality assessment of laboratory performance in bacteriology in the Eastern Mediterranean Region, 2011-2019. *East. Mediterr. Health J.* 28, 856–862. doi:10.26719/emhj.22.096

Turner, P., Fox-Lewis, A., Shrestha, P., Dance, D. A. B., Wangrangsimakul, T., Cusack, T.-P., et al. (2019). Microbiology Investigation Criteria for Reporting Objectively (MICRO): a framework for the reporting and interpretation of clinical microbiology data. *BMC Med.* 17, 70. doi:10.1186/s12916-019-1301-1

Turnidge, J. D., Jorgensen, J. H., and Zimmer, B. A. (2023). Susceptibility test methods: general considerations. *Man. Clin. Microbiol.*, 1411–1419. doi:10.1128/9781683670438.MCM.ch73

Walter, J., Noll, I., Feig, M., Weiss, B., Claus, H., Werner, G., et al. (2017). Decline in the proportion of methicillin resistance among *Staphylococcus aureus* isolates from non-invasive samples and in outpatient settings, and changes in the co-resistance profiles: an analysis of data collected within the Antimicrobial Resistance Surveillance Network, Germany 2010 to 2015. *BMC Infect. Dis.* 17, 169–177. doi:10.1186/s12879-017-2271-6

Wattal, C., Oberoi, J., Goel, N., Datta, S., Raveendran, R., and Prasad, K. J. (2019). Experience of Indian association of medical microbiology external quality assurance scheme centre, New Delhi: Challenges and quality assessment of clinical microbiology laboratories. *Indian J. Med. Microbiol.* 37, 163–172. doi:10.4103/IJMM.IJMM_19_356

Wonglumsom, W., Leepiyasakulchai, C., and Tiyasuttipan, W. (2008). External quality assessment in isolation and identification of bacteria. *Agric. Nat. Resour.* 42, 219–224. Available at: https://li01.tci-thaijo.org/index.php/anres/article/view/244597 (Accessed March 1, 2024).

TYPE Original Research
PUBLISHED 27 March 2024
DOI 10.3389/fmolb.2024.1373114

Check for updates

OPEN ACCESS

EDITED BY
Matteo Becatti,
University of Firenze, Italy

REVIEWED BY
Mohammad Tahir Siddiqui,
Indian Institute of Technology Delhi, India
Irina Nakashidze,
Shota Rustaveli State University, Georgia

*CORRESPONDENCE
Udo Reischl,
✉ Udo.Reischl@klinik.uni-regensburg.de

RECEIVED 19 January 2024
ACCEPTED 28 February 2024
PUBLISHED 27 March 2024

CITATION
Kremser M, Weiss N, Kaufmann-Stoeck A, Vierbaum L, Kappler S, Schellenberg I, Hiergeist A, Fingerle V, Baier M and Reischl U (2024), Longitudinal analysis of 20 Years of external quality assurance schemes for PCR/NAAT-based bacterial genome detection in diagnostic testing.
Front. Mol. Biosci. 11:1373114.
doi: 10.3389/fmolb.2024.1373114

Longitudinal analysis of 20 Years of external quality assurance schemes for PCR/NAAT-based bacterial genome detection in diagnostic testing

Marcel Kremser[1], Nathalie Weiss[1], Anne Kaufmann-Stoeck[1], Laura Vierbaum[1], Silke Kappler[1], Ingo Schellenberg[1,2], Andreas Hiergeist[1,3], Volker Fingerle[1,4], Michael Baier[5] and Udo Reischl[1,3]*

[1]INSTAND e.V., Society for Promoting Quality Assurance in Medical Laboratories, Duesseldorf, Germany, [2]Institute of Bioanalytical Sciences (IBAS), Center of Life Sciences, Anhalt University of Applied Sciences, Bernburg, Germany, [3]Institute of Clinical Microbiology and Hygiene, University Hospital Regensburg, Regensburg, Germany, [4]Bavarian Health and Food Safety Authority, Oberschleißheim, Germany, [5]Institute of Medical Microbiology, University Hospital Jena, Jena, Germany

Background: Quality control (QC), quality assurance, and standardization are crucial for modern diagnostic testing in the field of medical microbiology. The need for efficient QC to ensure accurate laboratory results, treatment, and infection prevention has led to significant efforts in standardizing assay reagents and workflows. External quality assessment (EQA) schemes, like those offered by INSTAND, play a vital role in evaluating in-house and commercial routine diagnostic assays, regarded as mandatory by national and global guidelines. The recent impact of polymerase chain reaction/nucleic acid amplification technology (PCR/NAAT) assays in medical microbiology requires that high-performing assays be distinguished from inadequately performing ones, especially those made by inexperienced suppliers.

Objectives: The study assesses the evolving diagnostic performance trends over 2 decades for the detection of EHEC/STEC, *Borrelia* (*B.) burgdorferi*, and MRSA/cMRSA. It explores the historical context of assay utilization, participant engagement, and rates of correct results in EQA schemes. The research seeks to identify patterns in assay preferences, participant proficiency, and the challenges encountered in detecting emerging variants or clinical strains.

Results: The study highlights the decline in in-house PCR assay usage, the emergence of new diagnostic challenges, and educational aspects within EQA schemes. Specific examples, such as the inclusion, in certain EQA surveys, of EHEC strains carrying *stx*-2f or *B. miyamotoi*, highlight the role of EQAs in increasing awareness and diagnostic capabilities. Advancements in MRSA detection, especially through the adoption of commercial assays, demonstrate the impact that technology evolution has had on diagnostic performance.

Conclusion: Achieving excellence in diagnostic molecular microbiology involves a multifaceted approach, including well-evaluated assays, careful instrumentation selection, and structured training programs. EQA schemes contribute significantly to this pursuit by providing insights into the evolving diagnostic landscape and identifying areas for improvement in the diagnostic workflow as well as in PCR/NAAT assay design.

KEYWORDS

EQA, *Borrelia burgdorferi*, MRSA, INSTAND, diagnostic molecular pathology

1 Introduction

Quality control (QC), quality assurance and standardization are among the most important prerequisites for modern diagnostic testing in medical microbiology and infectious diseases. Next to the use of well-evaluated assay concepts, the establishment and maintenance of efficient QCs are vital to ensuring the accuracy of laboratory results. This enables accurate patient identification and treatment as well as effective infection prevention (Badrick, 2021). Over the past decades, huge efforts have been made in standardizing assay reagents, creating diagnostic workflows, and incorporating internal controls with the aim of achieving results with the highest level of accuracy and reliability. External quality assessment (EQA) schemes are a crucial component in the reliable performance of routine diagnostic assays for pathogens or genetically encoded pathogenicity factors in medical microbiology and infectious diseases (Laudus et al., 2022). The value of regular participation is beyond dispute and hence mandatory in the official guidelines and regulations of most countries worldwide (De la Salle et al., 2017).

In the wake of the recent global pandemic, the commercial market has been flooded by many new assay concepts and instruments based on polymerase chain reaction/nucleic acid amplification technology (PCR/NAAT). These range from manual to semi- or fully automated systems and closed assay cartridges. Within this landscape, it is important to be able to distinguish between high-performing assays and assays from inexperienced suppliers that have inadequate analytical performance levels in real-world clinical settings.

Hence, there is a need to identify the many assays or test kits, supplied by inexperienced manufacturers, with inadequate performance in routine testing.

One of the significant challenges in diagnostic microbiology is the accurate detection of various pathogens, including bacteria and fungi. These microorganisms pose diverse challenges due to factors such as their genetic variability, rapid evolution, and the emergence of antimicrobial resistance or certain virulence factors. Accurate diagnosis is critical to reducing the spread of infectious diseases, optimizing patient management, and preventing adverse outcomes. Misdiagnosis or delayed diagnosis can lead to inappropriate treatment, disease progression, and potential transmission to others (Fournier et al., 2013). Therefore, the importance of precise and timely diagnosis cannot be overstated, especially in the context of these pathogens with significant nosocomial and/or public health implications.

INSTAND EQA schemes cover a broad range of relevant bacterial and fungal pathogens and are designed to identify and pinpoint potential weaknesses of certain PCR/NAAT assay concepts. Continuous participation not only serves as a benchmarking tool, as it is a way to obtain official certificates, it also has an educational effect. Retrospective studies reveal an improvement in laboratory performance among laboratories that regularly participate in EQA schemes, highlighting the educational role of EQAs (Keppens et al., 2018; Keppens et al. 2019; Keppens et al. 2021). The random inclusion of so-called "educative samples" among the selected target organisms reflects a primary commitment to the ongoing advances within the field of diagnostic medical microbiology and, consequently raises awareness of participants to new, emerging, or interesting genetic variants or clinical strains.

The INSTAND EQA project for the detection of bacterial DNA started in 2003 with biannual distributions of sample sets for *Chlamydia (C.) trachomatis*, *Neisseria gonorrhoeae*, *Bordetella pertussis*, *Helicobacter pylori*, enterohemorrhagic *Escherichia (E.) coli*/shigatoxigenic *Escherichia coli* (EHEC/STEC), *B. burgdorferi*, *Legionella pneumoniae*, *Salmonella enterica* and *Listeria* species (spp.). However, with the widespread adoption of PCR/NAAT-based assays in diagnostic medical microbiology, the EQA program has progressively broadened its spectrum and continues to grow.

The expanded EQA scheme now includes surveys for Methicillin-resistant *Staphylococcus aureus*/community acquired Methicillin-resistant *S. aureus* (MRSA/cMRSA), *C. pneumoniae*, *Mycoplasma pneumoniae*, *Bacillus anthracis*, *Coxiella burnetii*, *Francisella tularensis*, *Brucella* spp., Carbapenemases genes, toxinogenic *Clostridium difficile*, Vancomycin-resistant *Enterococci* (VRE), *Pneumocystis jirovecii*, and a comprehensive panel of bacterial urogenital pathogens that address recent multiplex PCR assay concepts.

Each EQA set comprises four samples containing various concentrations of the target organism as well as related species or *E. coli* cells as negative set members. Despite the great diagnostic potential of PCR testing, the success of each of its analytical applications is highly dependent on the reliability of the clinical samples containing nucleic acids for amplification. While EQA schemes may not perfectly mimic the range of different PCR inhibitors that are complicating real-world sample analysis (e.g., false-negative results or insufficient lower limits of detection) (Vesper et al., 2007), the proprietary matrix of lyophilized samples, composed of proteins, salts, and a significant number of human cells, enables the semiquantitative detection of human gene segments. This makes them valuable for use as purification, extraction, and/or inhibition controls.

While advancements in diagnostic technologies have undoubtedly improved the accuracy and efficiency of pathogen detection, there remain gaps in our understanding of the evolving trends in diagnostic methods and their performance over time. Existing literature highlights the transition from traditional culture-based methods to advanced molecular techniques like PCR/NAAT for rapid microbial identification and specific characterization (Weile and Knabbe, 2009; Das et al., 2017). However, there is limited research investigating the longitudinal trends in diagnostic accuracy and performance of these molecular assays, especially concerning their adaptation to changing clinical needs and emerging infectious threats. The study is the first to address this question by performing a longitudinal analysis over 20 years of EQAs for PCR/NAAT-based bacterial genome detection of EHEC/STEC, *B. burgdorferi* and for MRSA/cMRSA. Through this analysis, the study seeks to provide accessory insights into the evolving landscape of diagnostic testing, identifying patterns, challenges, and improvements in performance, thereby contributing valuable knowledge to the field.

2 Materials and methods

2.1 EQA procedure

The INSTAND EQA schemes for bacterial genome detection of EHEC/STEC (EQA 534), *B. burgdorferi* (EQA 535), and MRSA or cMRSA (EQA 539) were conducted globally twice a year (surveys in May and November) and contained four different samples per survey (4 × 0.3 mL). Detailed sample properties and compositions of microorganisms can be found in Supplementary Table S1. The stability and homogeneity of the EQA samples were assessed according to DIN EN ISO/IEC 17043:2023 standards (ISO/IEC17043:2023, 2023). To process the samples, the laboratories had to centrifuge the vials containing lyophilized material. The material was then reconstituted in 300 µL of sterile water (PCR grade) and incubated at room temperature for 20 min on an orbital shaker and/or with occasional vortexing. This resulted in suspensions comparable to native clinical specimens. 100 µL aliquots had to be processed using typical protocols for DNA extraction and PCR/NAAT assays established in the laboratories' routine diagnostic setting. Participating laboratories were tasked with determining qualitative outcomes (positive, negative, questionable) and were asked to submit their findings to the INSTAND "RV-Online" web portal (http://rv-online.instandev.de). Alongside the qualitative results, participants had to submit information on the methods used for DNA extraction and amplification, and specified which commercial kits were used or whether an in-house PCR assay (lab-developed test, LDT) was used. For all three EQA schemes, successful certification required an accurate determination of three out of four samples, as stipulated by the current guidelines of the German Medical Association (RiliBÄK) (Bundesärztekammer, 2023).

2.2 Data analysis and statistics

The EQA results for EHEC/STEC, *B. burgdorferi*, and MRSA or cMRSA were analyzed in a manufacturer-specific manner across surveys performed between November 2003 and May 2023. The MRSA EQA scheme started in November 2005. A limited number of results ($n = 2$) were reported in November 2007 for the EQA survey detecting EHEC/STEC, making a test-specific analysis statistically less robust. Hence, this survey was excluded from the study. This resulted in 39 surveys for EHEC/STEC, 40 for *B. burgdorferi*, and 36 for MRSA.

For all three pathogens, assay manufacturer collectives with the highest participant counts per survey were represented individually. In the case of EHEC/STEC, the six most common methods were presented, while for *B. burgdorferi* this number was seven, and for MRSA or cMRSA it was nine. The remaining commercial test kits or preconfigured PCR/NAAT assay concepts were combined into the category "other." Bar charts were used to illustrate the distribution of participating assay-specific laboratories over time for EHEC/STEC, *B. burgdorferi*, and MRSA or cMRSA. In order to discern potential trends over the years, percentages of correct results per date and sample were graphically depicted for each EQA scheme, with symbols indicating specific events. These events included the utilization of clinical variants, very low concentrations, and possible cross-contamination. A sample was considered correct when the presence or absence of the target microorganism was detected accurately. We analyzed the data based on the percentage of correctly identified samples in each survey per sample. Basic statistical analyses were performed using JMP 17.0.0 from SAS Institute (Cary, North Carolina, USA).

Overlay images were created using the Gnu image manipulator software 2.10.34.

3 Results

This study evaluated the inter-laboratory detection quality for EHEC/STEC and *B. burgdorferi* from November 2003 to May 2023, and for MRSA/cMRSA from November 2005 to May 2023. In order to identify the evolving trends, we analyzed up to forty EQA surveys during this period, looking at the number of participating laboratories, assay distribution (Figure 1), and rates of correct results (Figure 2).

The number of EQA participants for EHEC/STEC, *B. burgdorferi,* and MRSA increased from 30, 45, and 35, to 148, 131, and 331 respectively. At the beginning, 60% (MRSA), 87% (EHEC/STEC), and 93% (*B. burgdorferi*) of laboratories used in-house PCR assays. However, these percentages gradually declined over the years as commercially available assays gained prominence. By May 2023, the utilization of in-house PCR assays dropped to 24.8% (EHEC/STEC), 26.5% (*B. burgdorferi*), and 5.7% (MRSA) (Figure 1).

In order to analyze the progression of pass rates and testing quality, we graphically illustrated correct results [%] per date and sample, with symbols indicating either clinical variants, low pathogen concentrations, or cross-contamination (Figure 2). In the case of EHEC/STEC detection, correct results exceeded 85%, with instances of lower percentages typically corresponding to samples involving very low target organism concentrations or special clinical variants (Figure 2A). The only clinically relevant variants during the observation period were *stx*-2f and *eae* positive; the rates of correct results for these variants increased from about 24% to 60%.

Similar to the EHEC/STEC findings, the rate of correct results for *B. burgdorferi* consistently surpassed 90%, with instances of lower percentages often linked to very low pathogen concentrations, clinical variants, or possible cross-contamination (Figure 2B).

For the MRSA and cMRSA EQA schemes, instances with fewer correct results were notably associated with clinical variants and very low pathogen concentrations (Figure 2C). Additionally, green squares represent methicillin-susceptible *S. aureus* (MSSA) +

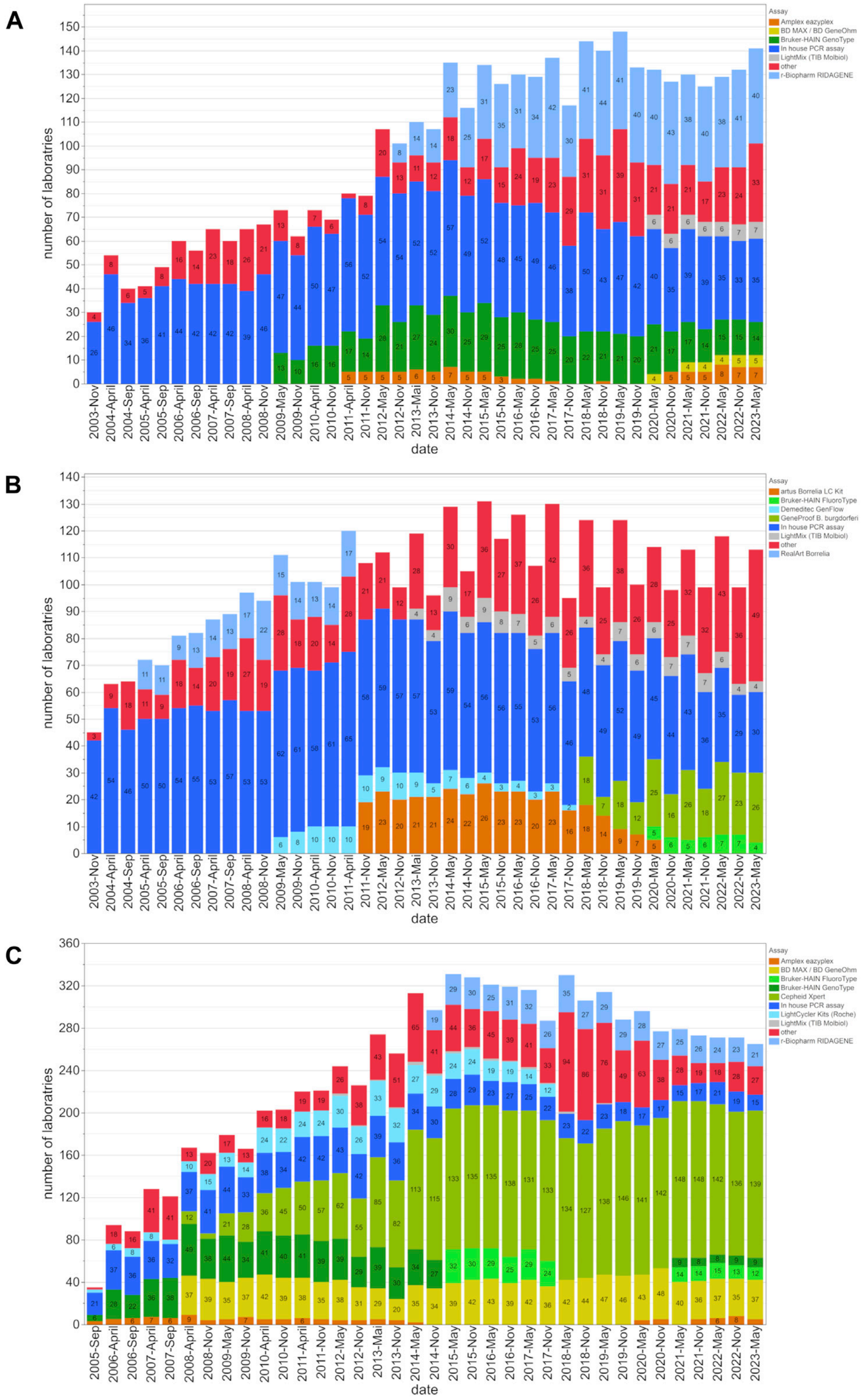


FIGURE 1
Assay distribution and number of participating laboratories from 2003 to 2023. This figure shows the distribution of assay utilization among participating laboratories and the changes in the utilization of these assays for the **(A)** EHEC/STEC, **(B)** *B. burgdorferi*, and **(C)** MRSA/cMRSA EQA schemes. The number of laboratories employing a certain assay type is indicated within the bars for each EQA scheme.

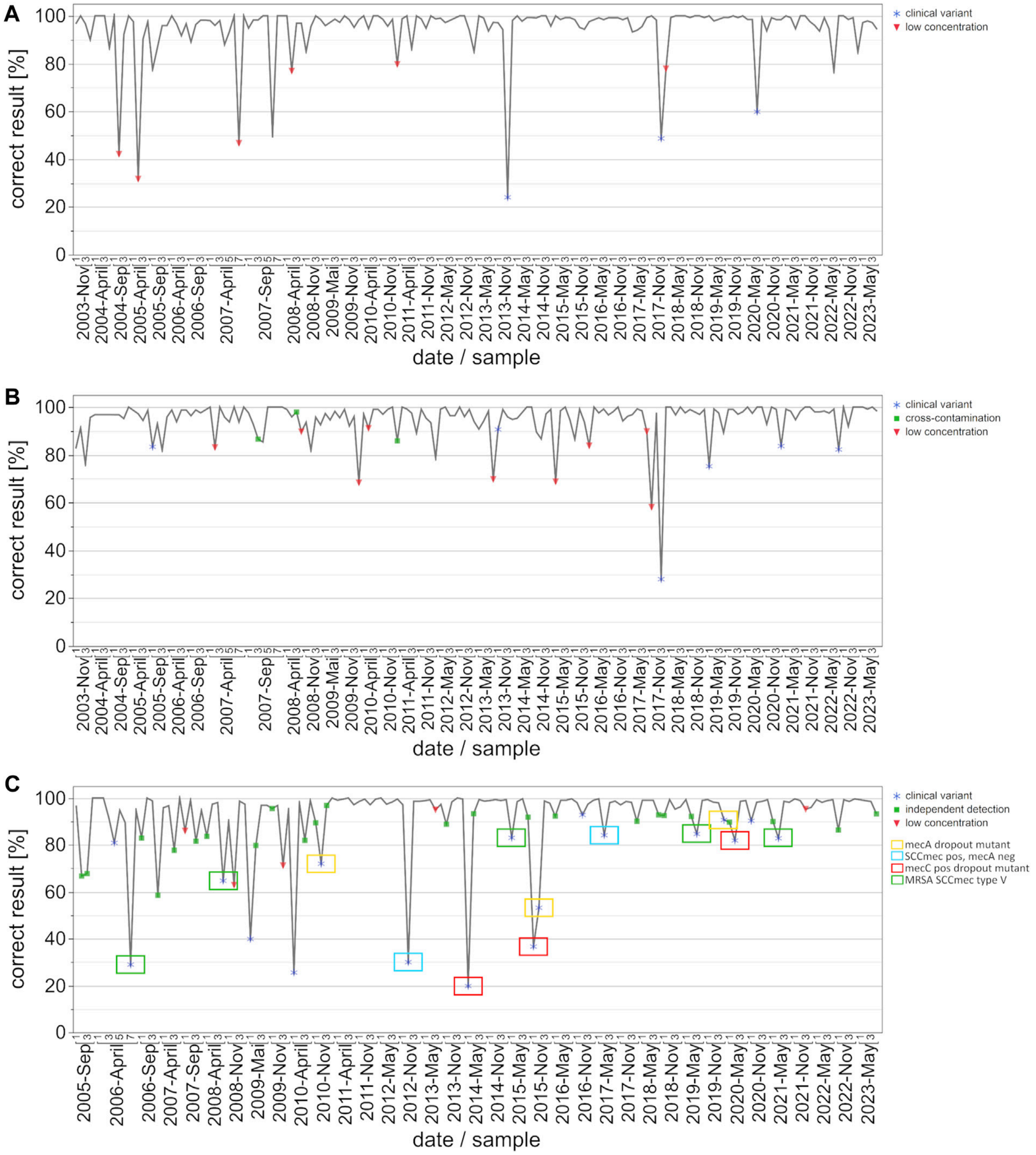


FIGURE 2
Development of correct results for the detection auf EHEC/STEC **(A)**, *B. burgdorferi* **(B)** and MRSA/cMRSA **(C)** from 2003 to 2023 with emphasis on key events/special sample composition. The data points on the graph represent the percentage of correct results per survey and date. Key events are defined as clinical variants (blue star), low concentrations of the respective target organisms (red triangle), and potential cross-contamination or independent detection (green square). The clinical variants were subdivided into four categories for MRSA: *mec*A dropout mutant (yellow box), SCC*mec* cassette positive but *mec*A negative MSSA (blue box), *mec*C positive MRSA variant (red box), and MRSA with an SCC*mec* type V cassette (green box).

coagulase-negative staphylococci (CONS) samples, for which "questionable" results from participants were included as correctly positive. This indicates the use of separate assays for detecting the mecA gene and a *S. aureus* species marker gene. Worth noting is the classification of clinical variants, with a particular focus on the four recurring categories: *mec*A dropout mutant (yellow box), SCC*mec* cassette positive but *mec*A negative MSSA (blue box), *mec*C positive MRSA variant (red box), and MRSA with an SCC*mec* type V cassette (green box). Rates of correct results for these four specific clinical variants improved over the years from under 50% to approximately 90%. The overall rate of correct results for MRSA and cMRSA consistently surpassed 90%, with instances of lower percentages often linked to very low pathogen concentrations, clinical variants, or the use of separate assays for detecting the mecA gene and a *S. aureus* species marker gene.

4 Discussion

Statistical analyses of EQA schemes, deliberately designed with highly diverse sample compositions in each survey, pose a complex challenge. These schemes lack a simple standard or comparator across different sample sets. Nonetheless, examining the percentage of correct results over nearly 2 decades offers valuable insights into assay standardization, coverage of variant bacterial pathogen strains, the analytical sensitivity for detecting relatively low concentrations of the respective target organisms, and the analytical specificity for distinguishing between the pathogen and the less pathogenic or apathogenic strains within a species or genus.

One illustrative example is the inclusion of clinically relevant variants of common pathogens like the Swedish *Chlamydia trachomatis* variant nvCT (Reischl et al., 2009). This new *C. trachomatis* variant was first identified in 2006 in the Swedish province of Halland and is characterized by a 377-bp deletion in the ORF-1 coding region of the multicopy cryptic plasmid. This region was targeted by both the Roche and Abbott *C. trachomatis* PCR assays available at the time. This nvCT strain was included in INSTAND's May 2009 EQA survey, in which the 128 participating laboratories used at least twelve different commercial PCR test kits or assays and a broad spectrum of in-house PCR assays. As expected, about 20% of the participants did not detect the *C. trachomatis* nvCT DNA in the sample when using the specific version of the Roche COBAS Amplicor CT/NG or several other, unspecified in-house PCR assays. When the nvCT strain was incorporated into a subsequent survey in May 2010, there was a notable increase in the accurate detection rate. It appears that the laboratories previously experiencing issues, as well as commercial PCR assay development teams, learned from this experience and subsequently redesigned their PCR assays to cover this variant strain (Reischl et al., 2010).

The examples of EHEC, *B. burgdorferi* and MRSA selected for this study emphasize the growing trend in utilizing prefabricated commercial PCR kits or closed cartridge-based PCR concepts (Figure 1). Many diagnostic laboratories still rely on established in-house or lab-developed tests (LDT) for the PCR/NAAT-based detection of the Shiga toxin-producing *E. coli* (EHEC/STEC). Although their prevalence slightly diminished around 2011 with the widespread availability of commercial PCR kits, in-house PCR assays, as shown in Figure 2A, continue to demonstrate high diagnostic accuracy and compete with commercial kits from various suppliers.

INSTAND's EQA scheme for the PCR/NAAT-based detection of EHEC/STEC (EQA 534) usually covers the various Shiga toxin genes and the putative accessory virulence marker genes of typical EHEC strains that occur around the world. The popular target genes include Shiga toxin gene variants *stx*-1, *stx*-1c, *stx*-2, *stx*-2c, *stx*-2d and *stx*-2e, as well as *eae*A (intimin) and E-*hly*A (enterohemolysin).

In 2000, a new Shiga toxin two variant (*stx*-2f) was identified in an *E. coli* strain isolated from pigeons (Schmidt et al., 2000). This observation enlarged the pool of *stx*-2 gene variants of human-pathogenic EHEC strains (Sonntag et al., 2005). It should be noted that the *stx*-2f encoding gene is quite distinct from other Shiga toxin gene variants at the nucleotide sequence level. This makes coverage by a common primer pair that targets conserved regions of *stx*-2, *stx*-2c, *stx*-2d, or *stx*-2e challenging. Consequently, modified or adapted assay designs require additional primer pairs and detection probes, complicating the composition of the PCR assay. Composition and subsequent comprehensive clinical re-validation of these assays may be needed.

The EHEC strain carrying *stx*-2f was first included in EQA 534 in November 2013. Similar to the situation with the aforementioned *C. trachomatis* variant, about 80% of the participants failed to detect the Shiga toxin gene variant in the sample when using various commercial test kits or other, unspecified in-house PCR assays. When the same strain was present in November 2017, the rate of correct detection increased to around 50% (58 out of 113 participants). By May 2020, this percentage had risen to around 60% (79 out of 132 participants), indicating the increased availability and use of re-designed commercial or in-house PCR assay concepts over the past decade. This situation is also nicely illustrated in the overall correct results depicted in Figure 2A, where the three outliers in November 2013 November 2017, and May 2020 correspond to the presence of EHEC strains carrying the *stx*-2f gene. Once again, this emphasizes the overall diagnostic advantages of incorporating such emerging or atypical strains of bacterial pathogens for educational purposes. It also raises awareness among colleagues in the fields of diagnostic microbiology and PCR/NAAT assay development of the rise of Shiga toxin variants in the EHEC circulating in animal and human populations. Moreover, the constellation depicted here represents similar situations in other INSTAND EQA schemes for PCR/NAAT-based detection of bacterial or fungal pathogens.

The PCR/NAAT-based detection of *B. burgdorferi* DNA is historically based on a variety of LDTs which evolved as robust and reliable diagnostic tools in the hands of experienced laboratories. With the increasing awareness of borreliosis as an emerging disease, several commercial kits have entered the market, supporting routine laboratories in expanding their diagnostic spectrum for detecting *B. burgdorferi* DNA in various types of clinical samples. Throughout the observed and analyzed time period, both in-house and commercial PCR assays consistently yielded high percentages of correct results, with only occasional interruptions due to samples containing very low numbers of target organisms (Figure 2B).

The *B. burgdorferi* PCR proficiency testing panel is designed for the specific and sensitive detection of *B. burgdorferi* sensu lato (s.l.) DNA, but the positive samples do not necessarily contain suspensions of "prototype" isolates of *B. burgdorferi* sensu stricto. Over the past 2 decades, many EQA surveys contained other *B. burgdorferi* genospecies or related species in individual samples. At least 21 different species are known to belong to the *B. burgdorferi* s.l. complex, which naturally present genetic differences in commonly used target genes. As part of our *B. burgdorferi* scheme, the May 2015 survey contained, in addition to three samples positive for the *B. burgdorferi* s.l. species, one sample with *B. miyamotoi* to challenge analytical specificities of PCR/NAAT assays used in the field. This species was first described in Japan in 1994. It belongs to the relapsing fever group of spirochetes but is transmitted by the same Ixodes ticks as *B. burgdorferi* s.l. in the United States, Asia and Europe. The *B. miyamotoi* sample was classified as false-positive by 36 of the 128 participating laboratories when certain commercial test kits or in-house PCR assays were used. A similar situation was observed in one sample of *B. hispanica* in the November 2020 survey. *B. hispanica* is not a member of the *B. burgdorferi* s.l. complex, but like *B. duttonii*, it is one of the causative agents for tick-borne relapsing fever that is present mainly in Spain and Northern Africa. This species is still extremely rare in Europe and of particular diagnostic importance for travelers with febrile illnesses. While the remaining 3 *B. burgdorferi* s.l. positive or negative samples in this particular survey were almost all correctly reported by the 98 participating laboratories, about 15% reported a false-positive result for *B. hispanica* organisms (Reischl et al., 2021). When sample sets contain analytical challenges in good faith and with an educative purpose, it is common practice in the supplementary documentation to encourage participants who obtained false-negative or false-positive results to re-evaluate their assay's analytical specificity and/or sensitivity. All in all, the inclusion of educative samples in conjunction with a corresponding scientific discussion is very well received by the participants.

MRSA detection improved significantly over the 20-year period with the broader introduction of commercial PCR assays and kits (primarily based on the detection of SCC*mec* cassettes) around the year 2010 (Figure 2C). The ability to discriminate between *mec*A-positive coagulase-negative staphylococcal species, *mec*A-negative *S. aureus* (MSSA), and the most critical *mec*A-positive *S. aureus* strains (MRSA) by covering the SCC*mec* cassette as an additional target is considered a milestone in rapid and reliable screening for MRSA in nasal swabs or other clinical specimens.

A second wave of improvement came with the awareness of *mec*C positive MRSA variants and their subsequent inclusion in some PCR assay concepts in 2017. Since then, an increasing number of commercial or in-house PCR concepts cover the *mec*C gene in addition to the *mec*A gene as potential methicillin-resistance markers in *S. aureus* organisms.

Furthermore, it should be noted that the EQA schemes use clinical isolates rather than classical type strains of a given species. This deliberate choice ensures a more representative assessment of diagnostic proficiency, as clinical isolates better reflect the complexities and variations encountered in real-world scenarios. By incorporating such clinically relevant strains, the EQA schemes aim to more accurately evaluate the ability of laboratories to detect MRSA or other pathogenic bacterial species of clinical relevance under conditions that closely mimic true clinical settings. Supplementary Table S2 provides additional insight into the diverse clinical variants considered in EQA scheme 539, specifically tailored to MRSA/cMRSA.

Over the past 2 decades, the percentage of in-house PCR assays has gradually decreased over the years. By May 2023, the use of in-house PCR assays decreased from 60% (MRSA), 87% (EHEC/STEC), and 93% (*B. burgdorferi*) to 24.8% (EHEC/STEC), 26.5% (*B. burgdorferi*), and 5.7% (MRSA). While the exact reasons for this shift remain unclear, a plausible explanation could be attributed to Regulation (EU) 2017/746 (IVDR) and its implementation of EU-wide, harmonized requirements for *in vitro* diagnostic medical devices in European healthcare institutions, which took full effect on 26 May 2022 (The European Parliament, 2017). Under the new EU regulation, healthcare institutions in the EU may continue to manufacture and use self-developed diagnostic products, provided they comply with the provisions outlined in Article 5 (5) of the regulation. However, certain requirements under the IVDR have been expanded beyond those of the previous regulations, resulting in increased validation and documentation efforts for medical laboratories (Hoffmuller et al., 2021). Consequently, the IVDR may be responsible for the gradual decline in the use of in-house PCR assays, as laboratories may increasingly switch to commercially available assays on the market that offer a more convenient solution amidst the increased regulatory requirements. In addition, the proliferation of commercial assays on the market provides laboratories with a wider range of options, further incentivizing the adoption of these commercially available assays over those developed in-house.

It is important to note that the overall diagnostic performance of individual laboratories is not solely determined by using "perfect" PCR assays. It also hinges on the careful selection and structured use of instrumentation, as well as the accurate execution of various manual steps throughout the entire workflow, including preanalytical and postanalytical processes.

Throughout the various EQA schemes, evident cross-contamination events during the consecutive steps of sample handling, automated or manual DNA preparation, and preparation of the PCR reaction mixtures mainly occurred when highly positive samples were present in individual sets. Laboratories that obtained such false-positive results due to contamination were clearly encouraged to monitor their individual diagnostic workflow and/or laboratory instrumentation for critical steps and initiate proper optimization measures. In addition to identifying general or specific shortcomings in the analytical sensitivity or specificity of individual PCR/NAAT assays, recognizing cross-contamination risks through regular participation in EQA schemes, and subsequently improving workflows contribute significantly to an overall enhancement of diagnostic quality.

Although this study provides valuable insights into longitudinal trends in diagnostic performance of PCR/NAAT-based bacterial genome detection, it is important to recognize several limitations. First, there may be potential bias in the selection of participants, as laboratories participate in EQA schemes on a voluntary basis, with participation being mandatory only for accredited labs, which may affect the representativeness of the data. In addition, variations in sample composition, including the concentration of target organisms and the presence of interfering substances, may affect assay performance and introduce bias into the results. Furthermore, it is important to note that our study utilized cultured samples rather than primary sample material. This distinction is particularly relevant since swabs often contain lower concentrations of target organisms compared to cultured samples.

Furthermore, the generalizability of our findings to broader scenarios beyond the specific infections analyzed needs to be considered. The dynamics of diagnostic performance observed in the EHEC/STEC, *B. burgdorferi* and MRSA/cMRSA assays may not be directly applicable to other pathogens or testing contexts. Therefore, caution should be exercised when extrapolating these results to other microbial targets or diagnostic settings.

Despite these limitations, our study underscores the importance of continued participation in EQA schemes and highlights the educational role of such programs in improving laboratory performance over time.

5 Conclusion

Achieving the highest level of performance in diagnostic molecular microbiology relies on a trifecta of critical elements (I) the use of well-evaluated PCR assay concepts or kits optimized with respect to analytical sensitivity and specificity, (II) a carefully selected and orchestrated instrumentation, and (III) structured programs for ongoing laboratory technician training to assure accurate execution of various manual steps within the workflow. Independent monitoring of the overall diagnostic performance is ultimately accomplished by regular participation in EQA schemes. Successfully meeting EQA requirements leads not only to essential certificates for maintaining the laboratory's official accreditation status but also to a better diagnostic efficiency that results in improved patient care.

In addition to assessing the diagnostic performance (analytical sensitivity and specificity) of different assays in individual laboratories, a statistical analysis of the results provides an actual snapshot of the technology and the use of commercial or in-house PCR/NAAT assays to detect a given pathogen among the broad and representative cohort of participants.

In essence, EQA schemes are not the sole solution but indeed one of the invaluable tools to preserving diagnostic quality. They provide early insights into potential shortcomings and weaknesses within the often complex and multifaceted diagnostic workflow, and contribute to the pursuit of excellence in diagnostic molecular microbiology.

Looking ahead, future research should continue to monitor diagnostic trends and performance to ensure the continued effectiveness of molecular microbiology diagnostics. In particular, efforts should be directed towards addressing continuous diagnostic challenges, such as the detection of new genetic variants as well as emerging antibiotic resistance genes or new putative virulence factors. In addition, expanding EQA schemes to include a wider range of pathogens and incorporating new technologies, such as next-generation sequencing, could further improve the quality and reliability of diagnostic tests. Collaboration between healthcare providers, regulators and industry stakeholders will be essential to drive innovation and improve patient outcomes in diagnostic medical microbiology.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://homepages.uni-regensburg.de/~reu24900/INSTAND_e.htm.

Author contributions

MK: Investigation, Visualization, Writing–original draft, Writing–review and editing. NW: Conceptualization, Writing–original draft, Writing–review and editing. AK-S: Writing–original draft, Writing–review and editing. LV: Writing–original draft, Writing–review and editing. SK: Writing–original draft, Writing–review and editing. IS: Supervision, Writing–original draft, Writing–review and editing. AH: Writing–original draft, Writing–review and editing. VF: Writing–original draft, Writing–review and editing. MB: Writing–original draft, Writing–review and editing. UR: Conceptualization, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2024.1373114/full#supplementary-material

References

Badrick, T. (2021). Integrating quality control and external quality assurance. *Clin. Biochem.* 95, 15–27. doi:10.1016/j.clinbiochem.2021.05.003

Bundesärztekammer (2023). Richtlinie der Bundesärztekammer zur Qualitätssicherung laboratoriumsmedizinischer Untersuchungen. *Dtsch. Ärzteblatt Jg.* 120, 21–22. doi:10.3238/arztebl.2023.rili_baek_QS_Labor

Das, S., Shibib, D. R., and Vernon, M. O. (2017). The new frontier of diagnostics: molecular assays and their role in infection prevention and control. *Am. J. Infect. Control* 45 (2), 158–169. doi:10.1016/j.ajic.2016.08.005

De la Salle, B., Meijer, P., Thomas, A., and Simundic, A. M. (2017). Special issue on external quality assessment in laboratory medicine - current challenges and future trends. *Biochem. Med. Zagreb.* 27 (1), 19–22. doi:10.11613/BM.2017.003

Fournier, P. E., Drancourt, M., Colson, P., Rolain, J. M., La Scola, B., and Raoult, D. (2013). Modern clinical microbiology: new challenges and solutions. *Nat. Rev. Microbiol.* 11 (8), 574–585. doi:10.1038/nrmicro3068

Hoffmuller, P., Bruggemann, M., Eggermann, T., Ghoreschi, K., Haase, D., Hofmann, J., et al. (2021). Advisory opinion of the AWMF *ad hoc* Commission *In-vitro* Diagnostic Medical Devices regarding *in-vitro* diagnostic medical devices manufactured and used only within health institutions established in the Union according to Regulation (EU) 2017/746 (IVDR). *Ger. Med. Sci.* 19, Doc08. doi:10.3205/000295

ISO/IEC17043:2023 (2023). *General requirements for the competence of proficiency testing providers*. Geneva, Switzerland: International Organization for Standardization.

Keppens, C., Boone, E., Gameiro, P., Tack, V., Moreau, E., Hodges, E., et al. (2021). Evaluation of a worldwide EQA scheme for complex clonality analysis of clinical lymphoproliferative cases demonstrates a learning effect. *Virchows Arch.* 479 (2), 365–376. doi:10.1007/s00428-021-03046-0

Keppens, C., Dufraing, K., van Krieken, H. J., Siebers, A. G., Kafatos, G., Lowe, K., et al. (2019). European follow-up of incorrect biomarker results for colorectal cancer demonstrates the importance of quality improvement projects. *Virchows Arch.* 475 (1), 25–37. doi:10.1007/s00428-019-02525-9

Keppens, C., Tack, V., Hart, N., Tembuyser, L., Ryska, A., Pauwels, P., et al. (2018). A stitch in time saves nine: external quality assessment rounds demonstrate improved quality of biomarker analysis in lung cancer. *Oncotarget* 9 (29), 20524–20538. doi:10.18632/oncotarget.24980

Laudus, N., Nijs, L., Nauwelaers, I., and Dequeker, E. M. C. (2022). The significance of external quality assessment schemes for molecular testing in clinical laboratories. *Cancers (Basel)* 14 (15), 3686. doi:10.3390/cancers14153686

Reischl, U., Ehrenschwender, M., Hiergeist, A., Maaß, M., Baier, M., Frangoulidis, D., et al. (2021). Bacterial and fungal genome detection PCR/NAT: comprehensive discussion of the November 2020 distribution for external quality assessment of nucleic acid-based protocols in diagnostic medical microbiology by INSTAND e.V. . *GMS Z Forder Qual. Med. Lab.* 12, Doc02. doi:10.3205/lab000042

Reischl, U., Schneider, W., Maaß, M., Straube, E., Fingerle, V., and Jacobs, E. (2010). Bakteriengenom-Nachweis PCR/NAT: auswertung des Ringversuchs April 2010 von INSTAND e.V. zur externen Qualitätskontrolle molekularbiologischer Nachweisverfahren in der bakteriologischen Diagnostik. *Der Mikrobiol.* 20, 193–209.

Reischl, U., Straube, E., and Unemo, M. (2009). The Swedish new variant of *Chlamydia trachomatis* (nvCT) remains undetected by many European laboratories as revealed in the recent PCR/NAT ring trial organised by INSTAND e.V., Germany. *Euro Surveill.* 14 (32), 19302. doi:10.2807/ese.14.32.19302-en

Schmidt, H., Scheef, J., Morabito, S., Caprioli, A., Wieler, L. H., and Karch, H. (2000). A new Shiga toxin 2 variant (Stx2f) from *Escherichia coli* isolated from pigeons. *Appl. Environ. Microbiol.* 66 (3), 1205–1208. doi:10.1128/aem.66.3.1205-1208.2000

Sonntag, A. K., Zenner, E., Karch, H., and Bielaszewska, M. (2005). Pigeons as a possible reservoir of Shiga toxin 2f-producing *Escherichia coli* pathogenic to humans. *Berl. Munch Tierarztl Wochenschr* 118 (11-12), 464–470. Available at: https://www.vetline.de/tauben-als-moegliches-reservoir-humanpathogener-shiga-toxin-2f-produzierender-escherichia-coli

The European Parliament (2017). *Regulation (EU) 2017/746 of the European parliament and of the council of 5 april 2017 on* in vitro *diagnostic medical devices and repealing directive 98/79/EC and commission decision 2010/227/EU*. Brussels, Belgium: Official Journal of the European Union. Available at: http://data.europa.eu/eli/reg/2017/746/oj/eng.

Vesper, H. W., Miller, W. G., and Myers, G. L. (2007). Reference materials and commutability. *Clin. Biochem. Rev.* 28 (4), 139–147. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2282402/

Weile, J., and Knabbe, C. (2009). Current applications and future trends of molecular diagnostics in clinical bacteriology. *Anal. Bioanal. Chem.* 394 (3), 731–742. doi:10.1007/s00216-009-2779-8

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 01 May 2024
DOI 10.3389/fmolb.2024.1338549

OPEN ACCESS

EDITED BY
Klaus-Peter Hunfeld,
Goethe University, Germany

REVIEWED BY
Surjendu Maity,
Duke University, United States
Flavia Martinello,
Federal University of Santa Catarina, Brazil

*CORRESPONDENCE
Lucy A. Perrone,
✉ lucy.perrone@ubc.ca

RECEIVED 14 November 2023
ACCEPTED 13 March 2024
PUBLISHED 01 May 2024

CITATION
Sreya M, Alam MS, Daula S, Lee C, Restelli V, Middlebrook K, Noble MA and Perrone LA (2024), Improving drinking water quality through proficiency testing—the impact of testing method and accreditation status on *Escherichia coli* detection by Canadian environmental testing laboratories.
Front. Mol. Biosci. 11:1338549.
doi: 10.3389/fmolb.2024.1338549

Improving drinking water quality through proficiency testing—the impact of testing method and accreditation status on *Escherichia coli* detection by Canadian environmental testing laboratories

Mahfuza Sreya[1], Md Saiful Alam[2], Sahibjot Daula[3], Caleb Lee[1], Veronica Restelli[1], Ken Middlebrook[4], Michael A. Noble[1] and Lucy A. Perrone[1]*

[1]Canadian Microbiology Proficiency Testing Program (CMPT), Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada, [2]School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada, [3]Department of Chemistry, University of British Columbia, Vancouver, BC, Canada, [4]Proficiency Testing Canada (PTC), Ottawa, ON, Canada

Water quality testing is crucial for protecting public health, especially considering the number of boil water advisories annually issued across Canada that impact daily life for residents in affected areas. To overcome these challenges, the development of drinking water safety plans and accessibility to regular testing using simple, rapid, and accurate materials are necessary. However, the significance of monitoring the accuracy of environmental microbiology testing laboratories cannot be overlooked. Participation in external quality assessment programs, such as those that include proficiency testing (PT), is a necessary risk management resource that ensures the effectiveness of these testing processes. Proficiency Testing Canada (PTC), in collaboration with the Canadian Microbiological Proficiency Testing (CMPT) program based at the University of British Columbia, have implemented a drinking-water microbiology PT program since 1996. Both PTC and CMPT are ISO/IEC 17043:2010-accredited EQA providers. The drinking water program provided PT challenges to subscribing testing laboratories twice per year. Each challenge consisted of four samples containing unknown concentrations of *Escherichia coli (E. coli)* and *Enterobacter* spp. Results from participants were assessed for accuracy based on the method of testing. This cross-sectional study evaluated 150 rural and metropolitan testing sites across Canada between 2016 and 2022. Multivariable logistic regression analysis was conducted to examine the impact of different testing methods and laboratory accreditation status on the proficiency scores. This approach enabled us to assess the association between multiple independent variables and the likelihood of achieving specific proficiency scores, providing insights into how testing methods and accreditation status affect overall performance. After adjusting for rural residence, testing time, and survey year, the membrane filtration method was positively associated with the likelihood of scoring satisfactory results compared to the enzyme-substrate method (OR: 1.75; CI:

1.37–2.24), as well as accreditation status (OR: 1.47; CI: 1.16–1.85). The potential for improvement in environmental laboratory testing performance through the implementation of regulated PT in drinking water safety plans is proposed, along with the need for reliable testing methods applicable to rapid drinking water microbiology testing.

KEYWORDS

drinking water, proficiency testing, accreditation, external quality assessment, microbial contamination

1 Introduction

Accurate quantification of microbial contamination in drinking water is imperative to the decisions made by Canadian provincial public health authorities for issuing public health alerts. Decisions ranging from bans on swimming in local lakes to issuing boil water advisories to entire communities are subject to reliable testing of microbial contamination indicators by regional environmental testing laboratories. Health Canada, which is responsible for issuing national health policy under the Government of Canada, has issued the *Guidance on Monitoring the Biological Stability of Drinking Water in Distribution Systems* (Health Canada, 2022) and *Guidance for providing safe drinking water in areas of federal jurisdiction* (Health Canada, 2021). Both documents outline recommendations for drinking water safety plans based on the World Health Organization's *Guidelines for Drinking-Water Quality* (WHO, 2017). Although these documents highlight the complexity of a sustainable drinking water monitoring system, both Canadian guidelines fail to emphasize the importance of enrolling in external quality assessment (EQA) as a means to verify the accuracy of the identification and quantification of microorganisms and to monitor the efficacy of the drinking-water safety plan after implementation.

One key pillar of EQA includes enrollment in a proficiency testing (PT) program, and which is also a requirement of laboratory accreditation to ISO/IEC 17025:2017 *General Requirements for the Competence of Testing and Calibration Laboratories*. PT programs send samples containing relevant hazard indicators of known concentration to testing laboratories for blinded analysis. These programs serve as one way to ensure the accuracy and reliability of detecting pathogenic indicators taken from regular drinking water samples, as well as the reporting and subsequent public health measures taken by the responsible laboratories (Molina-Castro et al., 2021). In particular, drinking water microbiology is often monitored by *Escherichia coli* (*E. coli*) detection. *E. coli* is an internationally recognized indicator of fecal contamination in drinking water that is associated with high public risk. *E. coli* is an ideal indicator as it is excreted in high numbers by animals and humans, tends to remain stable in drinking water, and is easily detected in comparison to enteric parasites and viruses (World Health Organization, 2017). Therefore, the presence of *E. coli* in drinking water is an indication of the possible presence of other disease-causing fecal microorganisms of concern. Such detection leads to the release of public health advisories, namely, boil water advisories.

In 2021, 18.4% of the total boil water advisories indicated "no applicable water quality reason" (Environment and Climate Change Canada, 2022). This suggests that reasons beyond the physical infrastructure of the drinking water system were cause for concern and led to the issuing of a public health notice. In Ontario, Canada, ISO/IEC 17025 accreditation and enrollment in PT became mandatory following the Report of the Walkerton Inquiry (O'Connor, 2002). In May 2000, 7 people died and over 2,300 became ill after agricultural runoff containing *E. coli* O157: H7 and *Campylobacter jejuni* entered the drinking water system in the small town of Walkerton, Ontario. An investigation into the incident found that the Walkerton Public Utilities Commission operators did not have the training and expertise to identify potential breaches in contamination, and budget cuts led to irregular monitoring of the drinking water safety system, which led to delayed boil water advisories in the region. Since then, the requirement for mandatory laboratory accreditation was fully integrated into the Ontario Safe Drinking Water Act in 2002 (Ontario, 2016). It was noted that although enrollment in accreditation and, thus, PT programs does not guarantee the accuracy of testing results, it offers a means of external and objective monitoring that has led to direct evaluation of laboratory testing and reporting systems.

Despite the definitive decisions made in Ontario after the Walkerton Inquiry, there is currently no legal requirement for other Canadian environmental testing laboratories to enroll in a PT program or obtain accreditation. The Standards Council of Canada (SCC) was established in 1970 to promote the voluntary adoption of standardized practices. Alongside the SCC, the Canadian Association for Laboratory Accreditation (CALA) and the Centre d'expertise en analyse environmentale du Québec in Quebec provide accreditation services to ISO/IEC 17025 and therefore require enrollment in PT. PT is regularly used internationally to evaluate the efficacy of drinking water testing within a region (Noble and Nikiforuk, 1996; Kelleher et al., 2017; Cao and Yang, 2020; Molina-Castro et al., 2021). Previous literature from CALA has also led to the evaluation of accredited and non-accredited environmental laboratories, where accredited laboratories were found to be more likely to show accurate results in PT (Morris and Macey, 2004; Middlebrook, 2017).

PT is a tool with which organizations can continually assess and monitor their testing process through interlaboratory comparison. This includes monitoring the training of laboratory staff, methodology, and the ability of staff to report accurate findings. This allows laboratories to identify risks within the drinking water quality system before public health notices are needed.

In 2006, PTC, in collaboration with the Canadian Microbiology Proficiency Testing program at the University of British Columbia, launched the "Microbiology in Water" program. Both PT providers

TABLE 1 An example of a Microbiology in Water survey set.

C05A-1 (mL)	C05A-2 (mL)	C05A-3 (mL)	C05A-4 (mL)
Escherichia coli: 30 CFU/100	*Escherichia coli*: 50 CFU/100	*Escherichia coli*: 20 CFU/100	*Escherichia coli*: 80 CFU/100
Enterobacter spp.: 30 CFU/100	*Enterobacter* spp.: 60 CFU/100	*Enterobacter* spp.: 70 CFU/100	*Enterobacter* spp.: 20 CFU/100

Samples are blinded to participants using the unique survey code. The CMPT senior technician selects the sample concentrations of each vial based on the sample concentrations of the previous year so as to avoid repeated survey challenges.

are accredited under ISO/IEC 17043:2010 *Conformity assessment—General requirements for the competence of proficiency testing accredited PT providers* (International Organization of Standardization, 2010). This program sends four samples of known concentrations of wild-type *E. coli* for blind analysis by the participating laboratory. These samples are combined with wild-type *Enterobacter* spp. at varying concentrations to emulate the lack of homogeneity in field drinking water samples. These samples are sent two times a year, in March and October, to monitor the ongoing performance of both public health and private testing laboratories that choose to enroll. This PT scheme is designed to be tested according to the environmental laboratory's typical operation.

Awareness of the efficacy of PT programs as a means of monitoring drinking water systems is lacking (Kelleher et al., 2017; Molina-Castro et al., 2021). Studies evaluating the efficacy of PT program implementation in drinking water do not mention the specificity and complexity of infectious pathogen testing. This study aims to evaluate the efficacy of common methods of *E. coli* quantification and accreditation status and their association with satisfactory proficiency scores in a nationwide study of environmental testing laboratories.

2 Materials and methods

2.1 EQA program design

PTC and CMPT collaborated in designing the "Microbiology in Water" PT program scheme. The scheme consisted of four different samples per survey. Each survey was sent biannually, typically in March and October of each year. Participants chose to subscribe to one survey or both throughout the year on a fee-for-service basis. Each sample consisted of 5 mL of bacterial stabilizer spiked with known concentrations of live *E. coli* and *Enterobacter* spp. to reflect the heterogeneity of fecal-contaminated field drinking water samples (Table 1). Samples were created and sent on the same day in order to maintain stability. PTC and CMPT recommended to participants that samples be tested within 96 h of the shipment date. Once received, participants diluted 1 mL of the sample into 1,000 mL of sterile distilled water (1:1,000 dilution), which could then be used to test for total coliforms and *E. coli* according to the participant's established protocols for membrane filtration (MF) using the agar(s) of choice, enzyme-substrate methods (EST), or the most probable number (MPN) method. Target concentrations of samples for each survey year were determined 1 year prior by the CMPT senior technician within a range of 20–100 CFU/100 mL per organism. This was done with careful consideration to avoid repeated testing values from the previous 2 years of surveys.

2.2 Sample preparation and validation

Each sample was prepared in single batches to maintain consistency across all survey sets. Bacterial stabilizer was prepared as previously described (Brodsky et al., 1978; Noble and Nikiforuk, 1996). Twenty-four hours before shipment of the survey, the colonies of *E. coli* and *Enterobacter spp.* were inoculated in respective tubes of 10 mL sterile Mueller Hinton broth (Oxoid, Nepean, Canada) and incubated at 37°C in O_2 overnight to obtain pure cultures of each organism. On the day of the shipment, the bacteria were centrifuged at 1711 *g* for 10 min at room temperature. The pellets were washed with sterile phosphate-buffered saline (PBS) twice, and resuspended in PBS. The initial quantity of each organism was determined by measuring the optical density at 570-nm. Serial dilutions were made in PBS to the previously established target concentration of each sample. The final dilution of each organism was pipetted into the beaker of bacterial stabilizer. Aliquots of 5 mL were dispensed into sterile vials (Sarstedt, Nümbrecht, Germany). The vials were sealed with o-ring caps, parafilmed, and shipped according to the UN3373 guidelines under the Transportation of Dangerous Goods Regulations (Transport Canada, 2022). 10% of samples from each batch were allocated for internal quality control at CMPT through systematic random sampling. Samples were divided and kept at room temperature and at 4°C to be tested at intervals of 24 h, 72 h, and 168 h after shipment. Sample concentrations were diluted as described above and validated for sample homogeneity and stability by MF.

2.3 Data collection and PT scoring

Participant results were collected electronically or by facsimile from participants. Data was collected on participant's location, accreditation status, testing method, date analyzed, and quantified *E. coli* count by CFU/100 mL or MPN/100 mL. The participating laboratories' names and addresses were deidentified. The data was stratified in Microsoft Excel by the given sample codes: C05A-1, C05A-2, C05A-3, and C05A-4. Then, each sample was stratified by method type to calculate the PT score of each survey. The PT score was calculated using the methods outlined in ISO/IEC 17043 (International Organization of Standardization, 2017). The scoring process consisted of calculating the mean, standard deviation, and z-score of the reported values by each testing method before assigning a PT score.

Outliers were defined as data points of extreme values and were identified visually (e.g., reported values of >1000 CFU/100 mL or MPN/100 mL). The decision to remove outliers visually was based

TABLE 2 ISO/IEC 17043:2010 Guidelines for proficiency test scores based on z-score.

Z-Score range	PT-score
$\|z\| < 2$	Satisfactory
$2 \le \|z\| < 3$	Questionable
$\|z\| \ge 3$	Unsatisfactory

on several careful considerations. Visually excluding data points at the extremities allowed us to directly assess the data in the context of the overall distribution of results and valid *E. coli* quantification as a proficiency testing provider. Stringent statistical cutoffs disproportionally removed valid data points that were critical to the nuances of each *E. coli* quantification method. Therefore, we were able to exercise discretion in considering not only the statistical anomalies, but also the relevance of each data point. We acknowledge that this approach can introduce subjectivity and should be used with discretion. To maintain rigor, we involved two independent researchers to score the proficiency testing survey and identify outliers in the process. The number of scored surveys was divided equally between the two researchers prior to analysis.

Starting with C05A-1, the mean (X) and the standard deviation (δ) of the reported values (x) for all testing methods were calculated using Excel. Each sample was given a z-score (z) based on the formula:

$$z = \frac{x - X}{\delta}$$

Each sample was then manually given a PT-Score based on the z-score criteria (Table 2). The mean, standard deviation, and z-score calculations were then repeated for each individual testing method, followed by manual PT-scoring of each method's reported values based on the z-scores for the C05A-2, C05A-3 and C05A-4 sample codes.

2.4 Statistical analysis

2.4.1 Variable selection

Statistical analysis in this study was conducted using STATA ver. 17 software (STATACorp, 2021). The distribution of each variable in relation to the outcome variable "Proficiency score" was compared using the Chi-square test, with results expressed as frequencies and percentages. To investigate the association between primary exposure "accreditation status" and "testing method" and the outcome of "proficiency score," we employed multiple logistic regression analysis for a complete case analysis. Relevant confounding variables, specifically rural residence, as well as risk factors for the outcome (testing time post-shipment and survey year), were included and conceptualized using a directed acyclic graph (DAG) (Supplementary Figure S1). The model was adjusted for the minimal sufficient adjustment set derived from the DAG, which included rural residence. Additionally, risk factors for the outcome, such as testing time post-shipment and survey year, were included in an automated backward stepwise regression with the Akaike Information Criterion (AIC) to enhance the precision of estimates.

TABLE 3 Characteristics of Canadian environment testing laboratories who were enrolled in external quality assessment (EQA) which includes proficiency testing (PT) program from 2016 to 2022 based on the total samples analyzed (N = 9,367).

Variable	Total samples analyzed (%)
Participant Characteristics	
Province	
Alberta	1,442 (15.4)
British Columbia	540 (5.8)
Manitoba	349 (3.7)
New Brunswick	662 (7.1)
Newfoundland and Labrador	214 (2.3)
Nova Scotia	694 (7.4)
Northwest Territories	51 (0.5)
Ontario	3,787 (40.4)
Prince Edward Island	55 (0.6)
Quebec	157 (1.7)
Saskatchewan	495 (5.3)
Yukon	77 (0.8)
Missing	844 (9.0)
Rural	
Yes	8,202 (87.6)
No	321 (3.4)
Missing	844 (9.0)
Testing Characteristics	
Accreditation (CALA)	
Yes	6,373 (68.0)
No	2,994 (32.0)
Testing Method	
Membrane filtration	5,850 (62.5)
Enzyme-substrate	2,653 (28.3)
Most probable number	813 (8.7)
Missing	51 (0.5)
Testing time post-shipment	
Less than or equal to 96 h	6,856 (73.2)
Greater than 96 h	2,513 (26.8)
Overall score	
Satisfactory	8,932 (95.4)
Unsatisfactory	435 (4.6)

2.4.2 Logistic regression modeling

Two model specifications were considered: the upper bound, which incorporated both the minimal sufficient adjustment set and the risk factors for the outcome, and the lower bound, which included only the

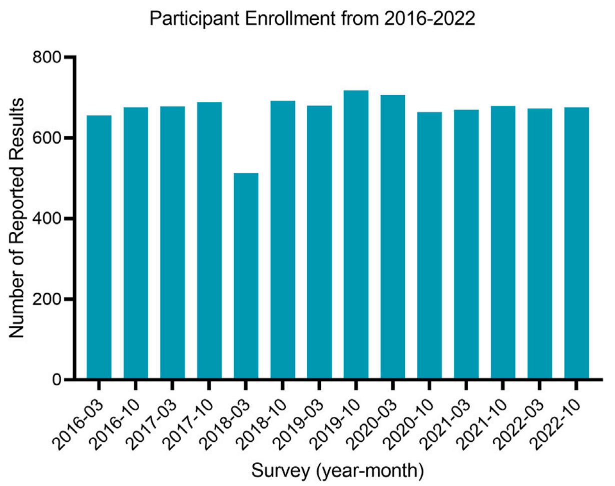


FIGURE 1
The number of reported results per survey in the "Microbiology in Water" proficiency testing program.

minimal sufficient adjustment set. Furthermore, we conducted tests to assess the interaction effect of rural residence in relation to exposure and outcome. Once the main effect model was defined, analysis of variance (ANOVA) was employed to evaluate whether the model that included the interaction terms was statistically significant. Model performance was assessed using a Receiver Operating Characteristics (ROC) curve, and goodness of fit was evaluated using the Hosmer-Lemeshow test. Measures of association were reported using odds ratios (OR) along with corresponding confidence intervals (CI). All statistical tests were two-sided, with a significance level set at $p < 0.05$.

3 Results

The majority of participants in the study (95.4%) achieved satisfactory proficiency testing (PT) scores, whereas a small proportion (4.6%) attained unsatisfactory PT scores, as presented in Table 3. Enrollment in the "Microbiology in Water" program had a total of 150 Canadian participants with consistent participation in surveys from 2016–2022, with a slight decrease in March of 2018 (Figure 1). The majority of participants were located in metropolitan areas (87.6%), with Ontario making up the highest proportion of participants (40.4%) and Alberta being the second highest (15.4%). The majority of participants reported results using membrane filtration (62.5%) or the enzyme-substrate method (28.3%) and were accredited through CALA (68.0%).

Satisfactory performance varied significantly between those using EST (93.8%), MF (96.5%), and MPN (93.6%) methods ($p <$ 0.001) and those accredited to ISO/IEC 17043 through CALA (96.2%) ($p <$ 0.001) (Table 4). After adjusting for confounding factors (testing time post-shipment, rural status, and survey year), the likelihood of satisfactory results is 1.75 times higher using the MF method compared to the EST method (CI: 1.37, 2.24; $p <$ 0.001), and 1.47 times higher when the laboratory is accredited (CI: 1.16, 1.85; $p <$ 0.001) compared to non-accredited laboratories (Table 5). In addition, there is a 75% decrease in the likelihood of a satisfactory result in participants that tested 96 h or longer post-shipment (%) (CI: 0.59, 0.98; $p <$ 0.05).

4 Discussion

As approaches such as drinking water safety plans become more widely accepted as a tool to develop a preventative framework in Canada, it is important to consider aspects such as testing methods and enrollment in accreditation, or PT, to improve drinking water quality within the context of individual communities. In this study, 150 Canadian laboratories participated in drinking water microbiology PT from 2016 to 2022, where the likelihood of a participant scoring a satisfactory result was significantly higher when using the MF method compared to the EST (Table 4), regardless of the culture medium used. MF itself proves to be the most prevalent method across Canada (Table 4), which further indicates its validity for reliable *E. coli* quantification. These findings are consistent with the recommendations of the *Standard Methods for the Examination of Water and Wastewater*, which establishes MF as a "gold standard" due to its ability to accurately quantify *E. coli* by colony count and high sensitivity for microorganisms within large volume samples (Rompré et al., 2002; Health Canada, 2020; American Public Health Association American Water Works Association Water Environment Federation, 2023).

The enzyme-substrate method has grown in popularity since its release in the 1990s as a rapid and streamlined MPN quantification method (Gorski et al., 2019). The EST method, namely, the Colilert tests (IDEXX Laboratories, Portland, ME, United States), utilizes the constitutive enzyme β-glucuronidase to detect *E. coli* by blue-white fluorescence due to its previous studies comparing standard MPN and MF methods to EST, which have produced variable results. Studies found that the EST tests generally underestimated *E. coli* recovery and produced a 10%–11% false negative rate (Schets et al., 2002; Fricker et al., 2010). Others comparing standard MPN and MF methods to EST have found that the methods show no significant difference in recovery of *E. coli* or have found EST to be a more sensitive method compared to the MPN method (Eckner, 1998; Kämpher et al., 2008); however, the PT samples used in this study control for confounding factors, such as non-coliforms in drinking water, that may lead to false-positive results.

The variety of EST findings, as indicated in this study, reflect the variability of the EST method as a whole. The accuracy of this can be significantly affected by the expression of β-glucuronidase, where recovery of 74 different fecal and environmental strains of *E. coli* has been found to be as low as 51.4% using the Colilert method (Maheux et al., 2008). In the context of this PT scheme for live *E. coli*, pre-analytical factors such as temperature fluctuations, pressure changes, and storage in a bacterial stabilizer could have affected the *E. coli* recovery by EST by the participants despite controlling for such factors through detailed consideration of the packaging.

Rurally located laboratory participants are more likely to use the EST method compared to the MF or MPN methods. Health Canada has outlined these concerns surrounding the EST method, yet rural communities face the challenge of a lack of trained staff, a lack of guidelines, or monitoring procedures specific to the rural laboratory (Health Canada, 2012; Lane et al., 2018). The EST method is efficient and requires minimal labour to process a sample. However, as communities begin to adopt water safety plans, careful consideration

TABLE 4 Bivariate analysis findings of Canadian environment testing laboratories who were enrolled in external quality assessment (EQA) which includes proficiency testing (PT) program from 2016 to 2022 stratified by satisfactory/unsatisfactory proficiency testing scores (N = 9,367).

	Proficiency testing score		
	Satisfactory (n = 8,932)	Questionable or unsatisfactory (n = 434)	p-value
Primary Outcomes			
Testing method			<0.001
EST	2,488 (93.8%)	164 (6.2%)	
MF	5,647 (96.5%)	203 (3.5%)	
MPN	761 (93.6%)	52 (6.4%)	
Accreditation status (CALA)			<0.001
Yes	6,129 (96.2%)	243 (3.8%)	
No	2,803 (93.6%)	191 (6.4%)	
Confounding factors			
Rural location status			<0.001
Yes	313 (97.5%)	8 (2.5%)	
No	7,884 (96.1%)	317 (3.9%)	
Testing time post shipment[a]			<0.001
<96 h	6,589 (96.2%)	263 (3.8%)	
≥96 h	2,240 (93.4%)	158 (6.6%)	
Survey (year)			0.069
2016	1,249 (93.8%)	83 (6.2%)	
2017	1,304 (95.5%)	62 (4.5%)	
2018	1,149 (95.4%)	55 (4.6%)	
2019	1,329 (95.1%)	68 (4.9%)	
2020	1,312 (95.8%)	58 (4.2%)	
2021	1,305 (96.7%)	44 (3.3%)	
2022	1,284 (95.3%)	64 (4.7%)	

[a]Testing time post-shipment n = 9,251.

must be given to the risks of using EST, as environmental strains may produce false negative results.

Our study revealed that participants that were accredited to ISO 17025:2017 were more likely to produce a satisfactory score (Table 4). These results are similar to those found in previous studies of performance among accredited laboratories (Morris and Macey, 2004). This may be a reflection of the benefits of ongoing monitoring of the quality management system of the individual laboratory, as regular enrollment in programs such as PT offers the chance for laboratories to reflect and improve not only the testing methods but the overall organization and documentation. However, enrollment in PT on its own can be beneficial. It is an established method for the QA/QC of safe drinking water systems that allows environmental laboratories to demonstrate the capability of accurately quantifying *E. coli* and demonstrate root cause analysis (Root et al., 2014).

The primary strength of this study lies in its robust data source, using a cross-sectional design that spans 6 years of participation. This comprehensive approach involves 150 participating laboratories form various Canadian provinces, enhancing the study's breadth and representativeness. The participant data was collected in real-time as reported by the participants and PT score allocation was blinded to avoid risk of investigator bias from prior knowledge of the laboratory. The PT score analysis was conducted using ISO/IEC 17043 guidelines in order to reflect the interlaboratory comparison that would be conducted by a PT or accreditation provider. Moreover, the data includes environmental laboratories from all 13 provinces and territories across Canada, which increases the generalizability of the water microbiology results and observations nationwide. However, the data used in this study was not initially conducted for the purpose of analysis beyond each survey. Therefore, there may be additional

TABLE 5 Crude and adjusted associations between the testing method, accreditation status and satisfactory/unsatisfactory (reference category) proficiency testing scores among Canadian environment testing laboratories who were enrolled in external quality assessment (EQA) which includes proficiency testing (PT) program from 2016–2022.

	Unadjusted odds ratio (95% CI)	Adjusted odds ratio (95% CI)
Testing Method		
EST	1.0 (Ref)	1.0 (Ref)
MF	1.84 (1.49, 2.28)***	1.75 (1.37, 2.24)***
MPN	0.97 (0.70, 1.33)	0.85 (0.58, 1.23)
Accreditation status to ISO/IEC 17025		
No	1.0 (Ref)	1.0 (Ref)
Yes	1.60 (1.31, 1.95)***	1.47 (1.16, 1.85)***
Rural location status		
No	1.0 (Ref)	1.0 (Ref)
Yes	1.54 (0.76, 3.14)	1.85 (0.89, 3.82)
Testing time post shipment		
<96 h	1.0 (Ref)	1.0 (Ref)
≥96 h	0.22 (0.18, 0.26)***	0.76 (0.59, 0.98)*
Year of program participation		
2016	1.0 (Ref)	1.0 (Ref)
2017	1.48 (1.04, 2.11)*	1.78 (1.13, 2.80)*
2018	1.33 (0.93, 1.90)	1.62 (1.03, 2.56)*
2019	1.22 (0.88, 1.71)	1.17 (0.78, 1.74)
2020	1.43 (1.01, 2.02)*	1.46 (0.96, 2.22)
2021	1.87 (1.28, 2.72)**	1.53 (1.01, 2.32)*
2022	1.28 (0.91, 1.81)	1.05 (0.71, 1.54)

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. CI, confidence interval.

confounding variables, such as technician skills and training, variation among method protocols, and participants' reporting, that could not be controlled for. Basing the analysis on participant reported results may lead to low confidence, as method protocols within each subgroup may vary drastically and methods reported at MPN may indicate the use of EST instead. As well, in order to evaluate a true positive correlation between continued participation in PT or accreditation and a satisfactory test result, the study would need to be repeated to look at results from particular laboratories beyond 2016 to evaluate the performance of selected laboratories over time. However, the nature of PT programs in drinking water microbiology is currently voluntary; therefore, it is difficult to follow laboratories continuously.

5 Conclusion

The MF method is more likely to produce satisfactory results in comparison to the EST method for the measurement of *E. coli* in drinking water. The variability in EST methods detected through PT participation calls for the need to develop efficient and reliable methods to quantify *E. coli* that can be used by laboratories regardless of geographic location. PT itself offers laboratories an opportunity to improve their general standards of practice, testing methods, and quality management.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

MS: Conceptualization, Methodology, Formal analysis, Investigation, Writing–original draft. MSA: Methodology, Formal analysis, Visualization, Writing–original draft. SD: Data curation,

Formal analysis, Writing–original draft. CL: Investigation, Methodology, Writing–review and editing. VR: Methodology, Writing–review and editing. KM: Data curation, Resources, Writing–review and editing. MAN: Writing–review and editing. LAP: Conceptualization, Supervision, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Funding for this work was provided through a fee for service mechanism established at CMPT and PTC. LP receives support from the Donald B. Rix Family Foundation.

Acknowledgments

We thank Esther Kwok, Fion Yung, and Shadi Alami for EQA program coordination and technical assistance. We thank the British Columbia Centre for Disease Control for their collaboration in the early phases of the water EQA program.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2024.1338549/full#supplementary-material

References

American Public Health Association American Water Works Association Water Environment Federation (2023). *Standard methods for the examination of water and wastewater*. Washington: APHA Press.

Brodsky, M. H., Ciebin, B. W., and Schiemann, D. A. (1978). Simple bacterial preservation medium and its application to proficiency testing in water bacteriology. *Appl. Environ. Microbiol.* 35, 487–491. doi:10.1128/aem.35.3.487-491.1978

Cao, N., and Yang, J. (2020). Proficiency test for determination of chlorate in drinking water. *Wei Sheng Yen Chiu* 49, 630–634. doi:10.19813/j.cnki.weishengyanjiu.2020.04.020

Eckner, K. F. (1998). Comparison of membrane filtration and multiple-tube fermentation by the colilert and enterolert methods for detection of waterborne coliform bacteria, *Escherichia coli*, and enterococci used in drinking and bathing water quality monitoring in southern Sweden. *Appl. Environ. Microbiol.* 64, 3079–3083. doi:10.1128/AEM.64.8.3079-3083.1998

Environment and Climate Change Canada (2022). Canadian environmental sustainability indicators: boil water advisories. www.canada.ca/en/environment-climate-change/services/environmental-indicators/boil-wateradvisories.html (Accessed August 28, 2023).

Fricker, C. R., Warden, P. S., and Eldred, B. J. (2010). Understanding the cause of false negative beta-D-glucuronidase reactions in culture media containing fermentable carbohydrate. *Lett. Appl. Microbiol.* 50, 547–551. doi:10.1111/j.1472-765X.2010.02834.x

Gorski, L., Rivadeneira, P., and Cooley, M. B. (2019). New strategies for the enumeration of enteric pathogens in water. *Environ. Microbiol. Rep.* 11, 765–776. doi:10.1111/1758-2229.12786

Health Canada (2012). Guidance on monitoring the biological stability of drinking water in distribution systems. https://www.canada.ca/en/health-canada/services/publications/healthy-living/guidance-monitoring-biological-stability-drinking-water-distribution-systems.html (Accessed August 28, 2023).

Health Canada (2021). Guidance on monitoring the biological stability of drinking water in areas of federal jurisdiction. https://www.canada.ca/en/health-canada/services/publications/healthy-living/guidelines-canadian-drinking-water-quality-guideline-technical-document-escherichia-coli.html (Accessed August 28, 2023).

Health Canada (2022). Guidelines for Canadian drinking water quality: guideline technical document – *Escherichia coli*. https://www.canada.ca/en/health-canada/services/publications/healthy-living/guidelines-canadian-drinking-water-quality-guideline-technical-document-escherichia-coli.html (Accessed August 28, 2023).

International Organization for Standardization [ISO] (2010). *Conformity assessment — general requirements for proficiency testing.* [ISO/IEC 17043: 2010]. Geneva: International Organization for Standardization.

International Organization for Standardization [ISO] (2017). *General requirements for the competence of testing and calibration laboratories.* [ISO/IEC 17025:2017]. Geneva: International Organization for Standardization.

Kämpfer, P., Nienhüser, A., Packroff, G., Wernicke, F., Mehling, A., Nixdorf, K., et al. (2008). Molecular identification of coliform bacteria isolated from drinking water reservoirs with traditional methods and the Colilert-18 system. *Int. J. Hyg. Environ. Health* 211, 374–384. doi:10.1016/j.ijheh.2007.07.021

Kelleher, K., Wong, J., Leon-Vintro, L., and Currivan, L. (2017). International Rn-222 in drinking water interlaboratory comparison. *Appl. Radiat. Isot.* 126, 270–272. doi:10.1016/j.apradiso.2017.01.036

Lane, K., Stoddart, A. K., and Gagnon, G. A. (2018). Water safety plans as a tool for drinking water regulatory frameworks in Arctic communities. *Environ. Sci. Pollut. Res. Int.* 25, 32988–33000. doi:10.1007/s11356-017-9618-9

Maheux, A. F., Huppé, V., Boissinot, M., Picard, F. J., Bissonnette, L., Bernier, J.-L. T., et al. (2008). Analytical limits of four beta-glucuronidase and beta-galactosidase-based commercial culture methods used to detect *Escherichia coli* and total coliforms. *J. Microbiol. Methods* 75, 506–514. doi:10.1016/j.mimet.2008.08.001

Microsoft Corporation (2018). Microsoft Excel. Available at: https://office.microsoft.com/excel.

Middlebrook, K. (2017). Do accredited laboratories perform better in proficiency testing than non-accredited laboratories? *Accredit. Qual. Assur* 22, 111–117. doi:10.1007/s00769-017-1262-z

Molina-Castro, G., Venegas-Padilla, J., Molina-Marcia, J., Scarioni, L., and Calderon-Jimenez, B. (2021). Improving the quality control of drinking water in Nicaragua through proficiency testing in a metrological multilateral cooperation project. *Sci. Rep.* 11, 16853. doi:10.1038/s41598-021-96230-w

Morris, A., and Macey, D. (2004). Laboratory accreditation: proof of performance for environmental laboratories—2001 study. *Accredit. Qual. Assur* 9, 52–54. doi:10.1007/s00769-003-0736-3

Noble, M. A., and Nikiforuk, S. (1996). Variability in urine culture reporting by Canadian microbiology laboratories. *Can. J. Infect. Dis.* 7, 247–249. doi:10.1155/1996/238498

O'Connor, D. R. (2002). *Report of the Walkerton Inquiry: the events of may 2000 and related issues: a summary*. Toronto: Ontario Ministry of the Attorney General.

Rompré, A., Servais, P., Baudart, J., de-Roubin, M. R., and Laurent, P. (2002). Detection and enumeration of coliforms in drinking water: current methods and emerging approaches. *J. Microbiol. Methods* 49, 31–54. doi:10.1016/s0167-7012(01)00351-7

Root, P., Hunt, M., Fjeld, K., and Kundrat, L. (2014). Microbiological water methods: quality control measures for federal clean water Act and safe drinking water Act regulatory compliance. *J. AOAC Int.* 97, 567–572. doi:10.5740/jaoacint.13-262

Saxena, T., Kaushik, P., and Krishna Mohan, M. (2015). Prevalence of *E. coli* O157: H7 in water sources: an overview on associated diseases, outbreaks and detection methods. *Diagn Microbiol. Infect. Dis.* 82, 249–264. doi:10.1016/j.diagmicrobio.2015.03.015

Schets, F. M., Nobel, P. J., Strating, S., Mooijman, K. A., Engels, G. B., and Brouwer, A. (2002). EU Drinking Water Directive reference methods for enumeration of total coliforms and *Escherichia coli* compared with alternative methods. *Lett. Appl. Microbiol.* 34, 227–231. doi:10.1046/j.1472-765x.2002.01075.x

StataCorp (2021). *Stata statistical software: release 17*. College Station, TX: StataCorp LLC.

Transport Canada (2022). Shipping infectious substances. https://tc.canada.ca/en/dangerous-goods/safety-awareness-materials-faq/industry/shipping-infectious-substances (Accessed August 30, 2023).

World Health Organization (2017). *Guidelines for drinking-water quality*. Geneva: World Health Organization.

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 05 February 2024
DOI 10.3389/fmolb.2024.1327699

OPEN ACCESS

EDITED BY
Dirk Schlüter,
Hannover Medical School, Germany

REVIEWED BY
Bharathikumar Vellalore Maruthachalam,
Janssen Research and Development,
United States
Martin Kammel,
GBD Gesellschaft für Biotechnologische
Diagnostik mbH, Germany

*CORRESPONDENCE
Jeremy V. Camp,
✉ jeremy.camp@meduniwien.ac.at

RECEIVED 25 October 2023
ACCEPTED 03 January 2024
PUBLISHED 05 February 2024

CITATION
Camp JV, Puchhammer-Stöckl E, Aberle SW and Buchta C (2024), Virus sequencing performance during the SARS-CoV-2 pandemic: a retrospective analysis of data from multiple rounds of external quality assessment in Austria.
Front. Mol. Biosci. 11:1327699.
doi: 10.3389/fmolb.2024.1327699

Virus sequencing performance during the SARS-CoV-2 pandemic: a retrospective analysis of data from multiple rounds of external quality assessment in Austria

Jeremy V. Camp[1]*, Elisabeth Puchhammer-Stöckl[1], Stephan W. Aberle[1] and Christoph Buchta[2]

[1]Center for Virology, Medical University of Vienna, Vienna, Austria, [2]Austrian Association for Quality Assurance and Standardization of Medical and Diagnostic Tests (ÖQUASTA), Vienna, Austria

Introduction: A notable feature of the 2019 coronavirus disease (COVID-19) pandemic was the widespread use of whole genome sequencing (WGS) to monitor severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections. Countries around the world relied on sequencing and other forms of variant detection to perform contact tracing and monitor changes in the virus genome, in the hopes that epidemic waves caused by variants would be detected and managed earlier. As sequencing was encouraged and rewarded by the government in Austria, but represented a new technicque for many laboratories, we designed an external quality assessment (EQA) scheme to monitor the accuracy of WGS and assist laboratories in validating their methods.

Methods: We implemented SARS-CoV-2 WGS EQAs in Austria and report the results from 7 participants over 5 rounds from February 2021 until June 2023. The participants received sample material, sequenced genomes with routine methods, and provided the sequences as well as information about mutations and lineages. Participants were evaluated on the completeness and accuracy of the submitted sequence and the ability to analyze and interpret sequencing data.

Results: The results indicate that performance was excellent with few exceptions, and these exceptions showed improvement over time. We extend our findings to infer that most publicly available sequences are accurate within ≤1 nucleotide, somewhat randomly distributed through the genome.

Conclusion: WGS continues to be used for SARS-CoV-2 surveillance, and will likely be instrumental in future outbreak scenarios. We identified hurdles in building next-generation sequencing capacity in diagnostic laboratories. EQAs will help individual laboratories maintain high quality next-generation sequencing output, and strengthen variant monitoring and molecular epidemiology efforts.

KEYWORDS

SARS-CoV-2, next-generation sequencing, diagnostic laboratories, Austria, COVID-19 diagnostic testing

1 Introduction

Within 1 month of the first detection of the novel coronavirus in December 2019, the dissemination of the virus sequence, as well as a publicly available database of similar sequences, allowed the development of the first diagnostic tests based on RT-qPCR for SARS-CoV-2 (Corman et al., 2020). A month later, in February 2020, reference laboratories in Europe (Reusken et al., 2020) and elsewhere were prepared to detect the virus with validated protocols. The new tools and kits prepared and sold by commercial entities initially were allowed emergency use certification as *in vitro* diagnostic (IVD) tests, pending more rigorous performance analysis. However, as the basic techniques for virus genome detection were already widespread in diagnostic laboratories, it was shown through external quality assessments (EQA) around the world that initial overall performance was high, with relatively few false negative results and almost no false positive results (Görzer et al., 2020). EQA schemes on SARS-CoV-2 genome detection served to inform public health authorities about the quality of epidemic data generated by diagnostic laboratories, but, equally importantly, informed participants about their performance relative to other labs so that they may identify areas for improvement (Buchta et al., 2023a). This is a critical process when new assays or techniques are implemented, as seen during the COVID-19 pandemic when the performance of a given assay was unknown and their implementation unfamiliar to some diagnostic laboratories (Buchta et al., 2023b).

Beyond mass testing to identify virus-positive individuals, the COVID-19 pandemic provided additional challenges for public health, namely tracking the emergence of novel viral variants. This task largely fell on the same diagnostic laboratories performing genome detection assays. Some laboratories opted to rely on more familiar assay formats (RT-qPCR and melting curve assays) to identify specific mutations characteristic of new lineages. While the assay format was familiar, the correct interpretation of the results was complex and often led to ambiguous results (Camp et al., 2021; Buchta et al., 2022a). Alternatively, some laboratories rapidly implemented whole genome sequencing (WGS) to characterize the viruses from patient samples. This expanding sequencing capacity was realized in laboratories across the globe, particularly in resource-poor countries where sequencing may not have been previously available. The rapid expansion of sequencing capacity was facilitated, in part, by i) the development and reliability of a whole genome sequencing strategy (Quick et al., 2017), ii) the availability of affordable 2nd and 3rd generation sequencing devices, and iii) the development and availability of bioinformatic pipelines. Together, generating consensus sequences of a virus from next-generation sequencing (NGS) platforms became achievable for the average clinical laboratorian. Online sequence databases began accumulating sequences, and, as the virus changed, governments encouraged–or even required–diagnostic and reference laboratories to provide sequences to these databases to track variants.

The result was that, at of the time writing, over 16 million sequences of SARS-CoV-2 have been uploaded to the Global Initiative on Sharing All Influenza Data ("GISAID", a database that emerged during the COVID-19 pandemic as a repository of SARS-CoV-2 sequences (Shu and McCauley, 2017)). Such comprehensive genetic coverage of a single virus has never been achieved before, and the compendium of genetic information has driven detailed analyses of the global evolution of SARS-CoV-2. However, the quality and/or accuracy of individual sequences is unknown. In contrast to virus genome detection assays, there were relatively few commercial options for preparing samples for SARS-CoV-2 WGS (First NGS, 2020), and these approaches were empirically tested, which could assist laboratories in selecting and validating suitable sequencing approaches (Charre et al., 2020; Liu et al., 2021). However, benchmarking bioinformatics pipelines for NGS data remains an issue in multiple fields, even when wet lab techniques in clinical diagnostics are established (Angers-Loustau et al., 2018; SoRelle et al., 2020; de Vries et al., 2021; Krishnan et al., 2021). Well into the COVID-19 pandemic, it was clear to some experts that deficiencies existed due to a lack of familiarity/competency in the bioinformatics analyses required for the producing quality data from NGS platforms, and this was hindering the utility of genomic surveillance (Hanahoe et al., 2021; Hodcroft et al., 2021).

The use of NGS in diagnostics has become more established in other fields. For example, there exist CE-certified *in vitro* diagnostic (IVD) kits for preparing samples from patients to identify germline or somatic-mutation related diseases, and these can be used on CE-IVD NGS platforms (e.g., Illumina MiSeqDx or Ion Torrent Genexus Dx). More recently, FDA-approved and CE-IVD tests can process patient samples (tissues or liquid biopsies) to detect and diagnose cancer on NGS platforms, as part of personalized medicine approaches (Jennings et al., 2017). Although NGS has been widely adopted by microbiology laboratories to complement diagnosis, identify microbial resistance, or profile the microbiome, the process towards validation of these techniques for routine use in diagnostics is less clear (Rossen et al., 2018). Similarly, there has been limited use of sequencing in routine clinical virology prior to the SARS-CoV-2 epidemic, mostly Sanger-sequencing relatively short nucleotide sequences for the purposes of antiviral resistance testing (e.g., HIV or Hepatitis B virus), and mostly limited to expert reference laboratories. The performances of some of these assays/techniques had been evaluated via EQA (Germer et al., 2013; Lee et al., 2020; Parkin et al., 2020), but these studies show that many laboratories preferred alternative methods for genotyping that did not require sequencing. At the time of writing, we know of only one CE-IVD assay for the sequencing of SARS-CoV-2; this status was only recently achieved and only under emergency use only regulations (Illumina COVIDSeq).

The benefits of virus genomic surveillance were evident prior to the COVID-19 pandemic (Dudas et al., 2017; Faria et al., 2017; Oude Munnink et al., 2021), were implemented in various countries during the COVID-19 pandemic (Oude Munnink et al., 2020), and will become essential to monitoring and controlling future epidemics. As SARS-CoV-2 variant monitoring was seen as an integral part of managing the pandemic, and drove public health policy decisions, the Austrian government provided a bounty on sequences submitted to public databases. Therefore, monitoring the performance of laboratories reporting sequencing data based on NGS techniques and ensuring their accuracy is of high importance. Having previously designed EQA schemes for SARS-CoV-2 virome detection (Buchta et al., 2023c), we sought to design an EQA scheme to test the performance of laboratories performing whole genome

TABLE 1 SARS-CoV-2 whole genome sequencing EQA schemes in Austria.

Round	Date	Sample	GISAID EPI_ISL_#	Mean Ct	Lineage
1	Feb 2021	hCoV-19/Austria/CeMM2633/2021	934568	29.5	B.1.1.7
		hCoV-19/Austria/MUW_1375876/2021	1191133	31.1	B.1.1.7
		hCoV-19/Austria/CeMM3247/2021	1008244	27.1	B.1.351
		hCoV-19/Austria/MUW_1320744/2021	913069	29.6	B.1.177
		hCoV-19/Austria/MUW_1358160/2021	913078	30.3	B.1.258
2	July 2021	hCoV-19/Austria/MUW_9133135702/2021	3144944	22.0	B.1.1.7 + S:E484K
		hCoV-19/Austria/MUW_1413581/2021	3144945	23.0	B.1.1.318
		hCoV-19/Austria/MUW_1379219/2021	1191134	29.5	B.1.351
		hCoV-19/Austria/MUW_1420272/2021	3144946	23.8	B.1.617.2
		hCoV-19/Austria/MUW_204840628007/2021	3144947	25.6	P.1
3	Feb 2022	hCoV-19/Austria/CeMM21006/2021	7798629	23.6	B.1.617.2
		hCoV-19/Austria/CeMM21823/2021	9011257	26.6	BA.1.1
		hCoV-19/Austria/MUW_1481609/2022	n/a	24.6	BA.2
		hCoV-19/Austria/CeMM20996/2021	7798619	24.6	AY.34
4	Oct 2022	hCoV-19/Austria/MUW_1513521/2022	13328434	22.0	BA.2
		hCoV-19/Austria/MUW_1513519/2022[a]	13328433	23.0	BA.2
		hCoV-19/Austria/MUW_1511131p2/2022[b]	15982848	18.0	BA.5.3
		hCoV-19/Austria/MUW_1511131p2/2022[b]	15982848	20.0	BA.5.3
5	May 2023	hCoV-19/Austria/MUW_1567739/2023	16006120	25.0	BF.11.3
		hCoV-19/Austria/MUW_1572048/2023	n/a	23.0	BQ.1.1.49
		Negative (HeLa cell culture supernatant)	n/a	n/a	n/a
		hCoV-19/Austria/MUW1584021/2023	17062380	20.0	XBB.1.5.12
		hCoV-19/Austria/MUW_1586996/2023	17247178	25.0	DB.1

[a]Sample with minor variants.
[b]Duplicate samples at two dilutions.

sequencing. Additionally, analyzing concordance between laboratories could provide some insight into the accuracy of publicly available sequencing data. As EQA also serves the function of providing direct feedback to participants, we tracked performance of some laboratories over several rounds, to see if performance improved.

However, to our knowledge there have been only two published EQA schemes to evaluate consensus sequences generated by virus whole genome sequencing, and there were none at the time when we implemented our scheme (Lau et al., 2022; Wegner et al., 2022). The goal of this project was to implement an EQA scheme to assess laboratories performing whole genome sequencing of SARS-CoV-2 in Austria. We report the results of a SARS-CoV-2 WGS EQA scheme over five rounds from February 2021 until June 2023, analyzing performance in terms of sequence accuracy, and the ability to interpret viral genomic data. We note changes in performance over time, and we discuss changes in the EQA scheme over time to highlight the difficulties that we experienced in designing such a scheme, particularly in comparison to other published schemes.

2 Materials and methods

2.1 Sample preparation and scheme organization

The EQA schemes in Austria were administered by the Austrian Association for Quality Assurance and Standardization of Medical and Diagnostic Tests (ÖQUASTA), providing the technical infrastructure associated with coordinating participant enrollment, distribution of sample materials, and collecting results. The Center for Virology at the Medical University of Vienna provided expertise in selecting and validating test samples as well as analyzing the reported results.

The EQA schemes for SARS-CoV-2 whole genome sequencing occurred in February 2021 July 2021, February 2022 October 2022, and May 2023 (Table 1). Enrollments were confirmed when the participant provided information on i) sequencing protocol including reagents or kits and specific primer panels; ii) sequencing platform; and iii) basics of bioinformatics pipeline used in analysis. Participants were mailed a panel of 4–5 samples mostly derived from residual material (oropharyngeal swab) received as part of routine diagnostic testing or occasionally plaque-purified virus isolate (Vero cells), and therefore no specific ethical approval was required. Samples were prepared by dilution in physiological saline, or (for the 4th and 5th rounds) in RNAlater® (Thermo Scientific). SARS-CoV-2 sequences from the samples were characterized initially by the reference laboratory as part of routine surveillance. The prepared panel was quantified by RT-qPCR targeting the E gene, and sequenced again. To ensure homogeneity of the prepared sample panel, each sample was sequenced 2–3 times in total by the reference laboratory prior to shipping; and at least one of those was after mimicking extreme shipping conditions (stored for 1 week at room temperature) to assess the stability of the samples. At least one of these quality control sequencing runs was performed on an Illumina MiSeq, and one was performed on a MinION Mk1c for the 3rd, 4th, and 5th rounds.

Samples representing contemporary circulating variants were selected each round (Table 1). Specimens with high estimated viral genome copy number were preferred, as dilutions were required to prepare the material for distribution. Mostly, "non-challenging" samples were selected, however some "challenging" or educational samples were included in later rounds. In round four (total four samples), a sample with minor variants (>30%) was included to test participant interpretation. In round four, two samples were the same isolate at two different dilutions (mean *Ct* values of 18 and 20) to test within-lab reproducibility. In round five, a sample negative for SARS-CoV-2 genome was included to test participant quality control measures and cross-contamination.

2.2 Reporting results

Participants were provided information about sample preparation, and instructed to sequence the panel using normal/routine protocols. Reporting could be carried out using an online system, however, the participants were provided a report form to fill out and (e-)mail or fax. Although the report format changed slightly as the scheme evolved, the requested results were the same in each round, designed to test two main competencies:

i. The ability to generate an accurate sequence.
ii. The ability to manage and interpret sequence data.

2.3 Technical evaluation of sequence accuracy

For the first two rounds, for each sample, the participants were requested to provide all nucleotide differences in comparison to the reference strain (the NCBI Reference Sequence "Wuhan-Hu1", GenBank accession number NC_045512.2). For rounds 3, 4, and 5, the participants were requested to submit sequence results in fastn format. Completeness was the percent of the genome reported as a nucleotide (A/T/C/G) or ambiguous symbol (K/M/R/S/W/Y). Participants were not penalized for missing genetic data (N's). In order to reduce bias, the consensus sequence from all submitted sequences (or the reported inferred sequence in rounds 1 and 2) and at least two sequences generated by the reference laboratory was considered the "true" sequence. Sequence accuracy was assessed by determining the number of differences in the submitted sequence (or the reported inferred sequence in rounds 1 and 2) compared to this consensus sequence. Mutations, insertions, and deletions compared to the consensus sequence were counted, and the number of differences was the Accuracy Score, where a higher value is worse. For the purposes of evaluation (pass/fail), participants passed if fewer than six differences were found relative to the consensus.

2.4 Technical evaluation of sequence interpretation

The ability to manage and interpret sequence data was assessed in two ways, by asking participants to characterize each sample based on its sequence. Therefore, the following "interpretation" results were evaluated based on each individual (submitted) sequence, and not on the "true"/consensus sequence that was used for the Accuracy Score.

2.4.1 Lineage interpretation

For each sample, the participants were requested to provide a Pangolin lineage assignment for their sequence (O'Toole et al., 2021). The submitted lineages were evaluated based on correctness (Pass/Fail/Can be improved) based on an independent assessment of lineage from the submitted fastn file.

2.4.2 Mutation reporting

For the 3rd, 4th, and 5th rounds, the participants were required to report amino acid mutations (and indels) with respect to the reference sequence ("Wuhan-Hu-1", NC_045512) in the spike protein only. This was done mostly out of practicality, given the increasing number of mutations in the virus genome, but also to continue testing the ability of the participant to interpret sequence data. Similar to the Accuracy Score, a Self-Reported Mutation Score was recorded as the number of differences between the submitted substitutions in the inferred spike protein and the independently-determined substitutions.

2.5 Statistical analysis

We used descriptive statistics to describe laboratory performance with respect to four measured results: completeness, accuracy, self-reported mutation score, and pangolin lineage assignment. We note the number of laboratories passing all samples, as well as the number of samples successfully sequenced as a function of their virus load (estimated average C_t value). The participation was relatively low (minimum 3 participants/round, maximum 8), with participation varying haphazardly. Furthermore, no two laboratories reported using exactly the same protocol. Therefore, statistical power was too low to perform robust statistical comparisons (e.g., if there was a relationship between accuracy and specific platforms or sample preparation kits).

TABLE 2 Participation and methods in SARS-CoV-2 whole genome sequencing EQA schemes in Austria.

ID	Sample preparation (primers if known)	Platform	Bioinformatics	Rounds
A	AmpliSeq	Illumina MiniSeq	DRAGEN Illumina	2, 3, 4
B	QIAseq	Illumina NextSeq	QIAGEN cov2insight	1, 2, 3
C	AmpliSeq	Ion Torrent	Ion Torrent suite	2, 3, 4, 5
D	AmpliSeq	Ion Torrent	Ion Torrent suite	1, 2, 3
E	NEBNext (ARTIC)	Illumina NextSeq	[Not reported]	2, 3, 4
F	NEBNext (ARTIC)	Illumina MiSeq and MinION Mk1c	In house	1, 2, 3, 4, 5
G	NEBNext (ARTIC)	Illumina MiSeq	nf-core/Viralrecon	3, 5

3 Results

3.1 Participants, platforms, protocols, and pipelines

Nine laboratories were registered over the course of five ring tests, two of which performed Sanger sequencing of a partial sequence, and were not considered in this manuscript (n = 7 participants). Participation varied within a given round between three and seven participants per round, with the nine laboratories participating in at least two rounds (Table 2). Incomplete results were reported by some participants and are considered missing data here. For example, one participant only submitted fastn sequences in one of the three rounds where it was specifically requested. Enrollment in the EQA scheme was voluntary, and the identity of the labs kept anonymous for evaluation purposes. However, we note that none of the laboratories performing virus genome surveillance in an official, government-sanctioned capacity participated, and participants were comprised of varying laboratory types (medical diagnostic and nonmedical; established clinical laboratories and newly implemented SARS-CoV-2-dedicated laboratories).

Four laboratories used exclusively Illumina platforms (three for three rounds and one for two rounds), and two used exclusively Ion Torrent platforms (for four and two rounds); one participant used Illumina for three rounds, and an MinION Mk1c (Oxford Nanopore Technologies) for two rounds.

Sample preparation consisted of targeted amplification using tiled amplicon approaches for all participants in all rounds, with three laboratories using AmpliSeq (two Ion AmpliSeq, ThermoFisher; one AmpliSeq for Illumina, Illumina), two using QIAseq (QIAGEN), and three using NEBNext® chemistry (New England Biolabs) prepared kits (*N.B.* two laboratories switched sample preparation methods during the course of the ring tests). Only one laboratory used two platforms during the five rounds: in-house reagents for amplification using ARTIC network primers followed by either the Nextera XT (Illumina) tagmentation kit or the NEBNext® (New England Biolabs) ligation barcoding kits on the Illumina platform, or the native barcoding kit (EXP-NBD104) with v9.4.1 chemistry (Oxford Nanopore Technologies) for multiplexing and library prep on the MinION platform. Those using the NEBNext chemistry or in-house procedures reported using the ARTIC network primers (v3 or v4.1) (Quick, 2020; Tyson et al., 2020), and one reported using VarSkip short primer set (v1 and v2, New England Biolabs).

In terms of bioinformatics pipelines to generate the consensus sequences, both laboratories using Ion AmpliSeq (ThermoFisher) kits and Ion Torrent platforms used the Ion Torrent suite software for data analysis. Similarly, the participant using the AmpliSeq reagents for Illumina platforms used the Illumina DRAGEN software, and the participant using the QIAseq used unspecified QIAGEN software. One laboratory used the Viralrecon Nextflow pipeline (from nf-core, https://nf-co.re), and one laboratory used an in-house pipeline for Illumina (fastp, minimap2, samtools, and iVar) and porechop with medaka_consensus (Oxford Nanopore Technologies) for sequencing on the MinION platform.

3.2 Sequence accuracy and completeness

In total, 99 test results were submitted over the five rounds (15-25-28-16-15 per round). Three of these (round 5) were from a sample that contained no SARS-CoV-2, and one of the three participants submitted a partial genome from this sample. From the 96 remaining test results, genome completeness was estimated from 53 sample-results. Completeness could only be taken from Rounds 1 and 2 if the participant specifically reported it (two did in round 1, n = 7, but none did in round 2); fastn files were never reported by a single participant in rounds 3 and 4 (n = 14); and if the sample was not sequenced it was considered "missing" and not 0%. Nonetheless, the mean genome completeness was 95% (range 45%–100%, sd = 12%) (Figure 1A). Only six results had a completeness <95%, (range 45%–92%) and were from a single participant on two consecutive rounds (3rd and 4th) for samples with C_t values in a range from 23.6 to 31.3. Ignoring these outliers, the mean completeness was 99% (median 100%, s.d. = 1%).

Among the 96 possible test results for SARS-CoV-2-positive samples, 6 (6%) were "not detected" and came from five unique samples with mean E gene C_t values of 25 (n = 1 of three submitted results for that round), 26.6 (1 of 7), 29.5 (2 of 5), 30.3 (1 of 3), and 31.1 (2 of 3) (Figure 1B). Three of the undetected results came from one lab in two rounds (using the QIAseq protocol on an Illumina NextSeq) for samples with $C_t \geq 29.5$.

The sequence Accuracy Score could not be calculated from 17 sequence results (6 times when the sample was reported "virus not detected" or 11 times when the participant did not submit the

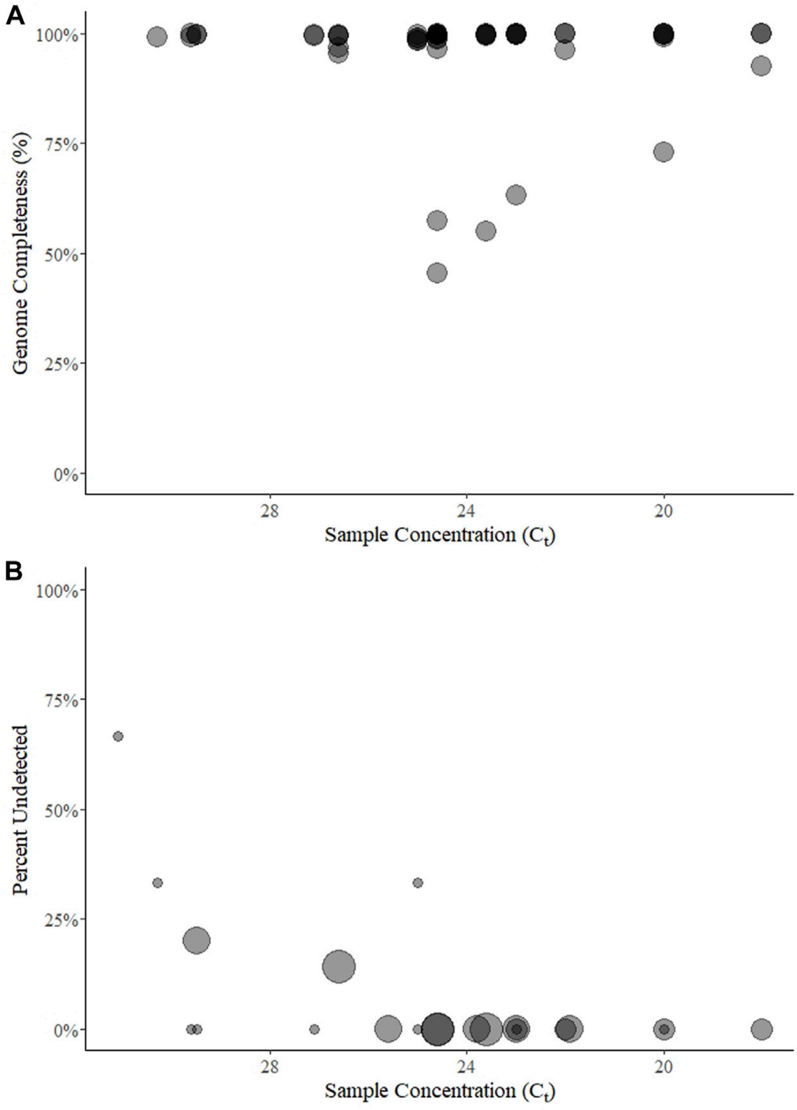


FIGURE 1
Genome completeness **(A)** and percent of samples undetected **(B)** over five rounds of a SARS-CoV-2 sequencing EQA in Austria. Points for genome completeness **(A)** are shown over approximate sample concentration (estimated by C_t value) from each submitted result (*n* = 54). The size of points in percent of results undetected **(B)** is relative to the number of participants (between 3 and 7) that submitted a result for a given sample (*n* = 22). In both panels the points are transparent gray and appear darker when overlapping.

requested results, as described above). From the remaining 79 test results, the mean Accuracy Score was 2.6 and the median was 1.0 (range 0–27, sd = 4.8) (Figure 2A). The distribution was skewed, with 48 (61%) results having one or fewer differences from the consensus and 36 results (45%) having no differences from the consensus. Only 11 results (14%) were categorized as failing (>5 differences), five of which came from one participant that participated in three rounds for a total of 13 tests (Supplementary Figure S1).

Over three rounds where accuracy could be independently validated (3rd–5th when fastn sequences were submitted), there were a total of 56 differences in submitted sequences compared to the corresponding consensus (Figure 3). Of these, 10 (18%) were ambiguous nucleotides (A/K/M/R/S/Y), and were not counted towards the Accuracy Score. Sequences with ambiguous nucleotide codes were all submitted by the same laboratory in the same round.

The remaining 46 differences comprised the Accuracy Scores, and could be classified into two main categories: frameshift mutations (indels, *n* = 12) and unique mutations (*n* = 34). As each submitted sample genome was compared to the consensus, the unique mutations could either be classified as “absent” (i.e., a mutation in the consensus relative to the reference strain that was *not* present in the submitted sequence) or “private” (i.e., a mutation present in the submitted sequence relative to the reference strain that was *not* present in the consensus).

There were twelve (of 46 = 26%) recorded instances where a participant submitted an indel in their sequence that would have resulted in a frameshift mutation. Eleven (92%) of the frameshifts were from a single lab in the fifth round that correctly identified

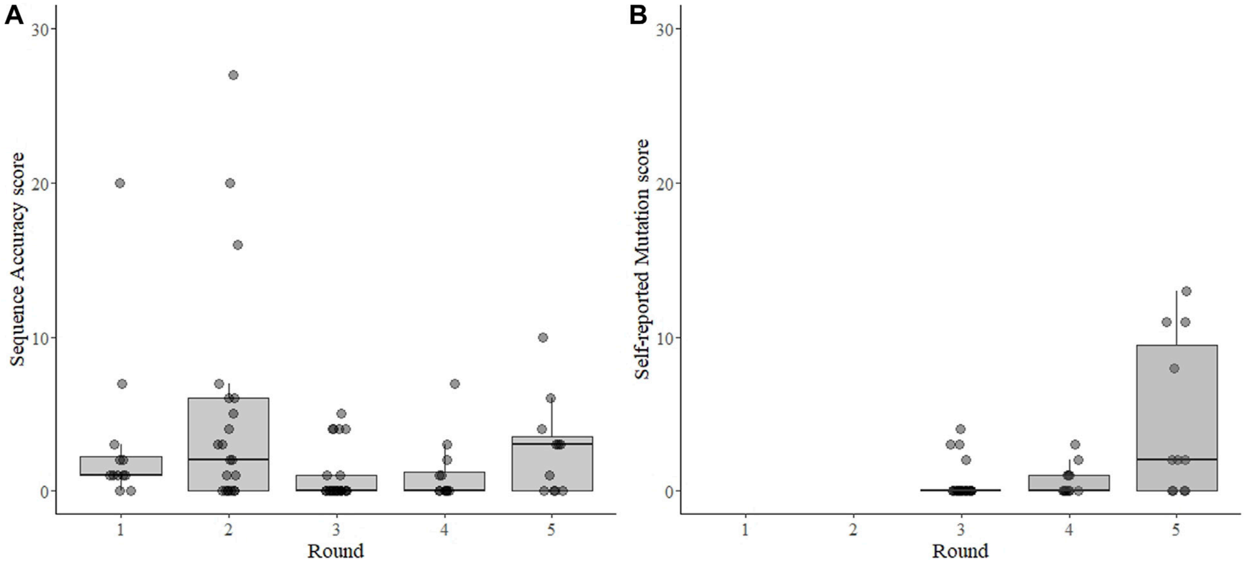


FIGURE 2
Sequence accuracy scores for SARS-CoV-2 sequencing EQAs **(A)** and scoring for self-reported mutation in the spike gene **(B)**. Each point represents the result from each sample submitted by each participant for each of five rounds of EQAs with boxes around the interquartile range and a thick horizontal line indicating the median value. The sequence accuracy score **(A)** is the number of differences from the consensus sequence of all laboratories for each sample. The accuracy score was independently verified for the 3rd, 4th, and 5th rounds, but were inferred from submitted mutations for the 1st and 2nd rounds as no fastn files were requested during those two rounds. The self-reported mutation score **(B)** is a count of the differences between the non-synonymous amino acid mutations in the spike protein reported by the participant and an independent analysis of the mutations by the evaluator, which was calculated for the 3rd, 4th, and 5th rounds. In both panels the points are transparent gray and appear darker when overlapping.

deleted regions, but reported them to be shorter than expected (e.g., a 9 nt deletion in ORF1a was reported as a 4 nt deletion). The majority (34/46 = 74%) of the errors were unique mutations, three of them were private mutations and 31 (67% of all differences) were "absent" mutations. Of note, 20 of the 56 differences (36%) were in the spike protein open reading frame (Figure 3).

One sample in round four contained minor variants at five sites at a level of 11%–37% of the called bases per site, as determined by the initial sequencing done by the reference laboratory (Supplementary Table). Three laboratories participated in this round, but one produced no sequence data for four of the five sites (Figure 3, "M"). The majority variant was detected by all laboratories at three sites, where the nucleotide composition was determined to be 63%, 85% and 89% of the reads. The minor variant was called by all laboratories except the reference laboratory once, where the minor variant was 34% of the reads; and all but one laboratory detected the majority variant at a site where the minor variant was 30%. Also in round four, two of the four laboratories reported exactly the same sequence for both replicated isolates (one even suggested in the optional notes that they were the same sample and not a cross-contamination) (Figure 3, "D"). One laboratory had different sequence scores and mutation scores for this sample (7 vs. 3, respectively for the C_t 18 the C_t 20 samples) (Figure 3, "4_A_3" and "4_A_4"). All of these were "absent" mutations found near regions with long stretches of Ns, as the participant reported 2223 and 8087 N's for samples, respectively (i.e., somewhat of a dilution effect). We could not verify the results of the fourth participant for these duplicated samples.

3.3 Sequence interpretation

The dataset to calculate the self-reported Mutation Score was nearly complete–missing only from the two samples in rounds 3-5 where the virus was "not detected", and not verifiable from one participant in rounds 3 and 4 that did not submit sequence files for the results. The mean mutation score was 1.5, with a median of 0.0 (range 0–13, s.d. 3.1) (Figure 2B). All four of the Mutation Scores higher than 4 were from a single participant in round 5 who apparently neglected to report mutations from the same region in each sample. The lineage was correctly reported 87 of 96 (91%) possible times. Of the 9 times it was incorrectly reported: 6 were from "not detected" results; two had too few data to assign a lineage, and one was from an apparent contamination.

4 Discussion

We analyzed the SARS-CoV-2 whole genome sequencing results from nine participants within Austria over five rounds of an EQA scheme. Our EQA scheme was designed to test two core competencies involved with SARS-CoV-2 sequencing: i) the ability to generate a consensus sequence from a sample and ii) the ability to interpret sequence data. The first competency deals with technical procedures involved with sample preparation (extraction, target enrichment/amplification, sequencing library preparation, analyzing raw sequencing data to prepare a consensus sequence). The second competency tests familiarity with sequence data: inspecting data, inferring coding regions and

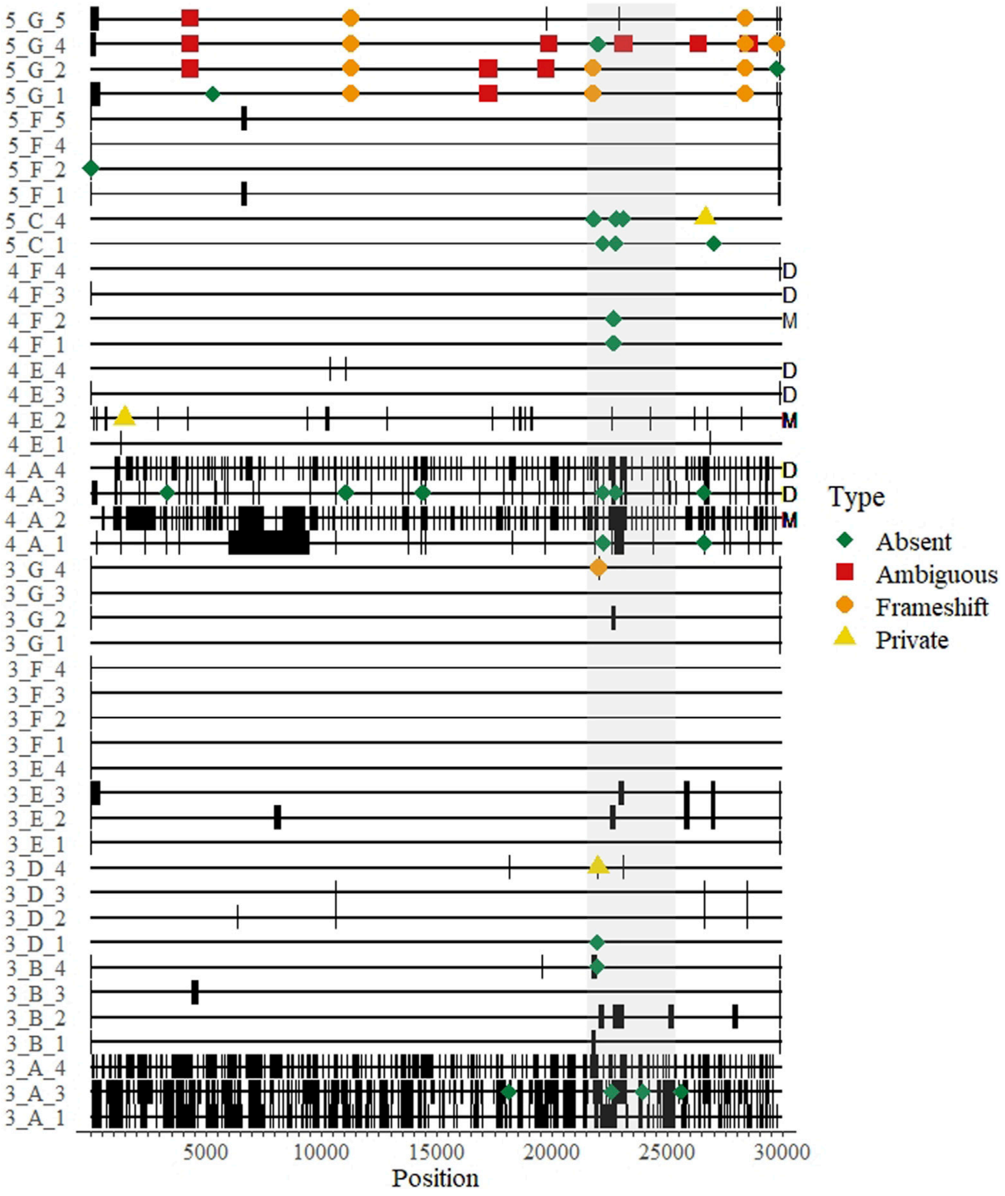


FIGURE 3
A genomic map of all whole genome sequences submitted over three rounds of a SARS-CoV-2 sequencing EQA in Austria missing data are thick black bands. Differences from the consensus and their approximate location on the genome are indicated by colored shapes: red squares indicate ambiguous nucleotides were called; orange circles represent indels that created a frameshift within an open reading frame; yellow triangles indicate a "private" difference from the consensus sequence that was not in the reference sequence (Wuhan-Hu-1, NC_045512); and green diamonds indicate mutation in the consensus sequence relative to the reference sequence (Wuhan-Hu1) that was not detected ("absent"). The sequence maps are grouped and labeled by round of EQA, laboratory letter as in Table 2, and Sample Number (Round_Lab_Sample) on the left. Letters on the right indicate results from duplicated samples ("D") or the sample with minor variants ("M").

identifying mutations, and utilizing new resources to categorize the sequence (identify a lineage).

Other EQA schemes for SARS-CoV-2 sequencing have focused on the first competency, and also noted that consensus genomes were highly reproducible across platforms (Wegner et al., 2022). Our "Accuracy" Score could be called a "reproducibility" scoring system, as we used the consensus of all results for a given sample as the "true" sequence. Therefore, our data provide an indication about the accuracy of consensus genomes submitted to public sequence databases. From our dataset, over half of the submitted sequences contained at least 1 difference from the consensus (Figure 2A). The ultimate source(s) of these differences were unclear, and mostly seemed stochastic, as they were distributed throughout the genome. Some errors tended to appear at the end of sequencing gaps, suggesting strict post-sequencing quality control and bioinformatics should implemented–this was particularly true of the laboratory receiving accuracy scores >1 for the duplicated sample in round four (Figure 3, "D"). In general, most of the differences counted in the score were "absent" mutations–i.e., the consensus was different from the reference strain at that site but a mutation was not detected by the participant. Furthermore, we included a sample with known minor variants in round 4, which we assume was the result of a coinfection (Figure 3, "M"). The results were mixed (Supplementary Table), suggesting differences could be introduced during sample preparation and not necessarily bioinformatics steps. Notably, the results from this sample might be informative to standardizing NGS techniques for identifying HIV anti-drug mutations (Lee et al., 2020; Parkin et al., 2020). However, in some cases the same errors were reproduced within a laboratory in the same round, suggesting they may be occasionally systematic, e.g., introduced by primers and/or related to the bioinformatics pipeline (Figure 3).

As a second metric of the ability to generate a consensus sequence, we reported the sequence completeness. Nearly every laboratory could produce a 99% complete genome, independent of sample C_t value within the range of ~20–31 C_t values (Figure 1). This highlights the usefulness of the tiled amplicon procedure for generating consensus genomes from patient samples (Quick et al., 2017; Tyson et al., 2020). All samples, particularly where coverage was less than 95%, showed a pattern of missing data consistent with amplicons generated for short-read sequencing (Figure 3). The fact that these errors were not associated with sample concentration indicates that these were due to errors in sample preparation (i.e., target enrichment steps or library preparation). Continued sequencing allows mutations that cause primer drop-out to be identified, and indeed tiled amplicon primer panels have undergone multiple versions as the SARS-CoV-2 virus has changed (Tyson et al., 2020). We noted some known primer dropout regions in the results associated with the ARTIC primers (Figure 3), but did not analyze whether other regions were associated with primer dropout, as we did not request participants submit raw sequence reads. As the target enrichment step relies on PCR, a key limitation of the approach is initial concentration of template in the test sample. However, we saw no relationship between genome completeness and estimated sample concentration (Figure 1A), and only a weak relationship between the percent of positive samples that were undetected per round and the estimated sample concentration (Figure 1B). We intentionally selected concentrated samples, and did not design a scheme to test the limit of detection. Others sequencing EQAs have included low concentration samples, and noted a significant reduction in the percent of completeness when sample concentration is diluted (Lau et al., 2022).

We observed additional, more serious, errors that were probably the result of pre-sequencing sample preparation steps. A participant in the first round - when variants contained fewer than 30 mutations per sample across the entire genome - reported mutations for two samples that were not included in the test panel, indicated laboratory contamination. Considering that most high throughput sequencing runs will include multiple SARS-CoV-2-positive samples, all with very similar genomes, cross-contamination remains difficult to detect. We tested this by including a negative sample in round 5 without informing the participants that one sample was negative. The sample was reported negative by two of the three participants. One participant stated that the sample "failed quality control" but submitted a partial sequence matching another sample in the panel and reported a lineage matching the same sample. It is crucial to maintain high standards to prevent and/or detect cross-contamination, particularly during post-PCR and pre-indexing library preparation steps.

The second competency that we evaluated was sequence interpretation. Viral variants can be detected and tracked with RT-qPCR techniques to identify SNPs based on melting curve analyses (Vogels et al., 2021). Interpreting these analyses requires additional competencies, as selecting the assays requires knowledge about circulating variants and familiarity with melting curve analyses (Camp et al., 2021; Buchta et al., 2022b). Whole genomes provide substantially more information and allow more specific identification of circulating variants. It was shown that laboratories that incorporated both RT-qPCR techniques and whole genome sequencing performed the best in terms of assigning lineages to a sample (Mögling et al., 2022). We found that identifying the lineage of a sequence was a relatively easy task for most laboratories that could generate a complete sequence from the sample. Similarly, "mistakes" in reporting mutations from the sequences could be explained by simple recording errors. Thus, in general, our observations indicated that errors in WGS results are likely from pre-sequencing sample preparation or post-sequencing consensus generation (bioinformatics). There were few errors in generating a lineage assignment or identifying mutations, regardless of the accuracy of the sequence.

Surprisingly, we saw that overall performance did not change drastically over time (Figure 4). The mean values of percent completeness, accuracy score, and mutation scores remained similar across all rounds, although there were fewer outliers in the later rounds (Figure 2). Considering individual laboratories, some did show improvement over the course of the rounds in which they participated. As the reported protocols did not change (except for one participant), it seems that competency and familiarity with the procedures involved in NGS increased over time. NGS requires some proficiency with bioinformatics, and this competency remains the critical hurdle for laboratories beginning to implement NGS. This issue was ostensibly solved early in the pandemic by the availability of many services to process raw NGS reads and produce a consensus SARS-CoV-2 genome (O'Toole et al., 2021; Cheng et al., 2023; Hadfield et al., 2018). Nonetheless, interpreting the data and maintaining good quality control of the output requires trained personnel. As bioinformatics tools continue to develop, maintaining their open-source nature allows laboratories across the globe access to similar and reproducible analysis pipelines.

At the time of the pandemic, there were few external quality assessment schemes designed for next-generation sequencing in clinical virology. Prior to the COVID-19 pandemic, a well discussed example was the development of an emerging EQA scheme using NGS for HIV drug resistance testing (Parkin et al., 2020), and the issues concerning reproducibility and accuracy when using NGS in clinical virology were already appreciated (Avila-Rios et al., 2020). NGS data are complex, and represent challenges for the laboratory as well as for designing quality assessment schemes. Evaluating the quality of these data can focus on many facets–genome detection, sample preparation and sequencing, and bioinformatics–each of which could be evaluated separately. Our EQA scheme focused on generating consensus sequences from primary material, and had to evolve to accommodate the progressive accumulation of mutations in the SARS-CoV-2 genome. Namely, we changed the requirements to submit a consensus sequence instead of submitting all mutations. In retrospect, requesting a consensus sequence was essential to allow

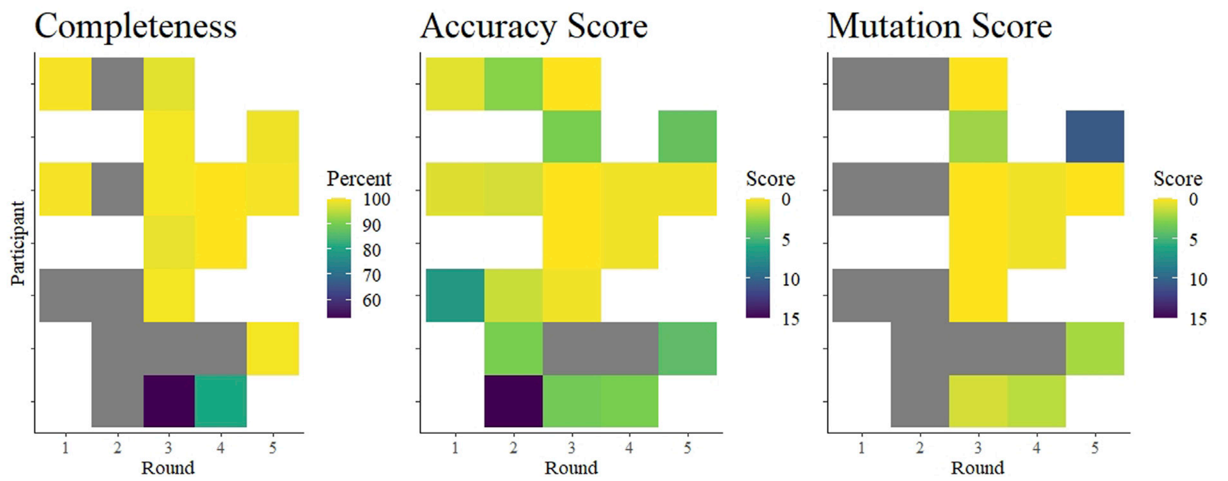


FIGURE 4
Summary of average completeness, accuracy score, and mutation score of each participant in each round. The colored tiles are shaded according to a gradient as indicated in each legend. White indicates no participation in that round. Gray squares indicate participation but data were missing. Specifically, accuracy scores could be interpreted from submitted data in rounds 1 and 2, but were not independently verifiable as no fastn sequences were submitted. Completeness scores from two laboratories in round 1 were based on submitted genome completeness information that was not specifically requested.

independent evaluation of the sequencing. Fundamentally, the consensus sequence represents the desired and reported outcome from WGS, similar to how detected/not detected is the fundamental reported outcome from virus genome detection assays. We did not request that participants provide quality metrics (e.g., average depth) or specifics about the analysis used to generate the consensus sequence, but these information would also be helpful in evaluating laboratory performance. However, such quality metrics are not requested commonly from virus detection EQA schemes (e.g., reporting specific Ct values, limit of detection, use of internal controls, etc.), and therefore we did not specifically request it. Others that requested participants report sequencing depth as part of their EQA scheme did not demonstrate any clear specific relationship between depth and other metrics (Lau et al., 2022; Wegner et al., 2022). Whether to request quality metrics from participants will depend on the fundamental goal of the EQA.

A principal limitation of this study is the fact that all laboratories did not participate in each round, and it is unknown whether registered participants intentionally did not participate because of "bad" results. Similarly, the small sample size prevents us from making comparisons between reagents, protocols and pipelines. However, these comparisons are probably better left to more controlled settings, where a single laboratory evaluates sample preparation methods across multiple samples (Charre et al., 2020; Liu et al., 2021) or bio-informatics pipelines with multiple datasets (Lee et al., 2020; SoRelle et al., 2020; de Vries et al., 2021; Krishnan et al., 2021). For viral WGS, EQA schemes are best suited to test two competencies, separately: sequencing and bioinformatics. By removing the bioinformatics portion, the ability to prepare a sample for sequencing and sequence the sample can be assessed while controlling for variability or deficiency in bioinformatics. Such deficiencies may be the reason for some of the seemingly random errors we observed (Figure 3). Similarly, a separate EQA scheme should test bioinformatics competency by supplying raw sequencing data and requesting a consensus sequence. This would eliminate the possibility of poor performance in sequence interpretation being due to problems with sequencing, and this is particularly important if sample degradation during shipping and handling is expected to be a source of error. The organization of bioinformatics schemes would rely on the availability of data from sequencing platforms in use and might only be feasible for ring tests with much larger enrollment. Such bioinformatics EQAs could also include evaluation of the applications of NGS–which we call "interpretation"–such as lineage assignment or inferring mutations. Others have also used EQA to measure applied competencies such as assigning samples to specific transmission clusters based on WGS data (Lau et al., 2022; Wegner et al., 2022). However, we observed very few errors in "interpretation" that could not be explained by simple data-entry/recording errors. Indeed, the most difficult aspect of organizing this EQA was designing the report form in a such a way that the requested data (fastn file, lineage assignment, and mutations) were correctly requested/reported. In the last round, there were still participants that could/did not follow the instructions, which we assume reflects the complexity of the analysis for the average clinical laboratorian.

Overall, we found that laboratories were prepared to implement next-generation sequencing methods to sequence whole genomes of SARS-CoV-2 relatively early in the COVID-19 pandemic. Excluding few outliers, participants achieved nearly 100% coverage of the genome, producing very nearly identical consensus sequences. Although we identified some deficiencies, we noted improvements within laboratories (Figure 4). As a caveat, these data may not be representative of the quality of sequences from Austria, as it is unknown whether the participants were performing routine sequencing with the reported protocols, or whether they were using the EQA to validate new protocols. Moreover, we know that not all Austrian laboratories submitting SARS-CoV-2 sequences to public databases participated in this EQA scheme.

Nonetheless, our data suggest that the learning curve for implementing next-generation sequencing in a diagnostic laboratory is steep, but surmountable, and EQAs can help by providing independent feedback. These competencies are applicable towards achieving increased monitoring of seasonal virus epidemics, as well as enabling readiness for monitoring a future emerging zoonotic virus.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical approval was not required for the studies involving humans because the samples included in the study were completely anonymized residual material intended for routine diagnostic testing by participant laboratories. The studies were conducted in accordance with the local legislation and institutional requirements. The human samples used in this study were acquired from a by-product of routine care or industry. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

JC: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Writing–original draft. EP-S: Resources, Supervision, Writing–review and editing. SA: Conceptualization, Methodology, Resources, Supervision, Writing–review and editing. CB: Conceptualization, Data curation, Methodology, Project administration, Resources, Supervision, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors wish to thank Christian Borsodi, Andreas Rohorzka and Thomas Urbanek at the Center for Virology for excellent technical assistance. We thank Irene Görzer at the Center for Virology for scientific advice and discussion. We thank Jovana Jovanovic and Dominik Kollros at ÖQUASTA for administrative and organizational assistance. We owe our sincere thanks to all participant laboratories, who we do not list here to maintain their anonymity.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2024.1327699/full#supplementary-material

References

Angers-Loustau, A., Petrillo, M., Bengtsson-Palme, J., Berendonk, T., Blais, B., Chan, K. G., et al. (2018). The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies. *F1000Res.* F1000Res, ISCB Comm J-459. doi:10.12688/f1000research.14509.2

Avila-Rios, S., Parkin, N., Swanstrom, R., Paredes, R., Shafer, R., Ji, H., et al. (2020). Next-generation sequencing for HIV drug resistance testing: laboratory, clinical, and implementation considerations. *Viruses* 12 (6), 617. doi:10.3390/v12060617

Buchta, C., Aberle, S. W., Allerberger, F., Benka, B., Gorzer, I., Griesmacher, A., et al. (2023b). Performance of SARS-CoV-2 nucleic acid amplification testing in Austria as measured by external quality assessment schemes during 3 years of the COVID-19 pandemic: an observational retrospective study. *Lancet Microbe* 4, e1015–e1023. doi:10.1016/S2666-5247(23)00286-0

Buchta, C., Camp, J. V., Jovanovic, J., Puchhammer-Stockl, E., Strassl, R., Muller, M. M., et al. (2022b). Results of a SARS-CoV-2 virus genome detection external quality assessment round focusing on sensitivity of assays and pooling of samples. *Clin. Chem. Lab. Med.* 60 (8), 1308–1312. doi:10.1515/cclm-2022-0263

Buchta, C., Camp, J. V., Jovanovic, J., Radler, U., Benka, B., Puchhammer-Stockl, E., et al. (2022a). Inadequate design of mutation detection panels prevents interpretation of variants of concern: results of an external quality assessment for SARS-CoV-2 variant detection. *Clin. Chem. Lab. Med.* 60 (2), 291–298. doi:10.1515/cclm-2021-0889

Buchta, C., Springer, D., Jovanovic, J., Borsodi, C., Weidner, L., Sareban, N., et al. (2023c). Three rounds of a national external quality assessment reveal a link between disharmonic anti-SARS-CoV-2 antibody quantifications and the infection stage. *Clin. Chem. Lab. Med.* 61 (7), 1349–1358. doi:10.1515/cclm-2022-1161

Buchta, C., Zeichhardt, H., Aberle, S. W., Camp, J. V., Gorzer, I., Weseslindtner, L., et al. (2023a). Design of external quality assessment schemes and definition of the roles of their providers in future epidemics. *Lancet Microbe* 4 (7), e552–e562. doi:10.1016/S2666-5247(23)00072-1

Camp, J. V., Buchta, C., Jovanovic, J., Puchhammer-Stockl, E., Benka, B., Griesmacher, A., et al. (2021). RT-PCR based SARS-CoV-2 variant screening assays require careful quality control. *J. Clin. Virol.* 141, 104905. doi:10.1016/j.jcv.2021.104905

Charre, C., Ginevra, C., Sabatier, M., Regue, H., Destras, G., Brun, S., et al. (2020). Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evol.* 6 (2), veaa075. doi:10.1093/ve/veaa075

Cheng, Y., Ji, C., Zhou, H. Y., Zheng, H., and Wu, A. (2023). Web resources for SARS-CoV-2 genomic database, annotation, analysis and variant tracking. *Viruses* 15 (5), 1158. doi:10.3390/v15051158

Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K., et al. (2020). Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill.* 25 (3), 2000045. doi:10.2807/1560-7917.ES.2020.25.3.2000045 de Vries, J. J. C., Brown, J. R., Fischer, N., Sidorov, I. A., Morfopoulou, S., Huang, J., et al. (2021). Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. *J. Clin. Virol.* 141, 104908. doi:10.1016/j.jcv.2021.104908

Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., and Rambaut, A. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* 544 (7650), 309–315. doi:10.1038/nature22040

Faria, N. R., Quick, J., Claro, I. M., Theze, J., de Jesus, J. G., and Pybus, O. G. (2017). Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* 546 (7658), 406–410. doi:10.1038/nature22401

First NGS (2020). First NGS-based COVID-19 diagnostic. *Nat. Biotechnol.* 38 (7), 777. doi:10.1038/s41587-020-0608-y

Germer, J. J., Abraham, P., Mandrekar, J. N., and Yao, J. D. (2013). Evaluation of the Abbott HBV RUO sequencing assay combined with laboratory-modified interpretive software. *J. Clin. Microbiol.* 51 (1), 95–100. doi:10.1128/JCM.02155-12

Görzer, I., Buchta, C., Chiba, P., Benka, B., Camp, J. V., Holzmann, H., et al. (2020). First results of a national external quality assessment scheme for the detection of SARS-CoV-2 genome sequences. *J. Clin. Virol.* 129, 104537. doi:10.1016/j.jcv.2020.104537

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., et al. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34 (23), 4121–4123. doi:10.1093/bioinformatics/bty407

Hanahoe, H., Austin, C. C., and Shanahan, H. (2021). Sharing COVID data? Check these recommendations and guidelines. *Nature* 592 (7855), 507. doi:10.1038/d41586-021-01028-5

Hodcroft, E. B., De Maio, N., Lanfear, R., MacCannell, D. R., Minh, B. Q., Schmidt, H. A., et al. (2021). Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature* 591 (7848), 30–33. doi:10.1038/d41586-021-00525-x

Jennings, L. J., Arcila, M. E., Corless, C., Kamel-Reid, S., Lubin, I. M., Pfeifer, J., et al. (2017). Guidelines for validation of next-generation sequencing-based oncology panels: a joint consensus recommendation of the association for molecular pathology and college of American pathologists. *J. Mol. Diagn* 19 (3), 341–365. doi:10.1016/j.jmoldx.2017.01.011

Krishnan, V., Utiramerur, S., Ng, Z., Datta, S., Snyder, M. P., and Ashley, E. A. (2021). Benchmarking workflows to assess performance and suitability of germline variant calling pipelines in clinical diagnostic assays. *BMC Bioinforma.* 22 (1), 85. doi:10.1186/s12859-020-03934-3

Lau, K. A., Horan, K., Goncalves da Silva, A., Kaufer, A., Theis, T., Ballard, S. A., et al. (2022). Proficiency testing for SARS-CoV-2 whole genome sequencing. *Pathology* 54 (5), 615–622. doi:10.1016/j.pathol.2022.04.002

Lee, E. R., Parkin, N., Jennings, C., Brumme, C. J., Enns, E., Casadella, M., et al. (2020). Performance comparison of next generation sequencing analysis pipelines for HIV-1 drug resistance testing. *Sci. Rep.* 10 (1), 1634. doi:10.1038/s41598-020-58544-z

Liu, T., Chen, Z., Chen, W., Chen, X., Hosseini, M., Yang, Z., et al. (2021). A benchmarking study of SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples. *iScience* 24 (8), 102892. doi:10.1016/j.isci.2021.102892

Mögling, R., Fischer, C., Stanoeva, K. R., Melidou, A., Almeida Campos, A. C., Drosten, C., et al. (2022). Sensitivity of detection and variant typing of SARS-CoV-2 in European laboratories. *J. Clin. Microbiol.* 60 (12), e0126122. doi:10.1128/jcm.01261-22

O'Toole, A., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J. T., et al. (2021). Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* 7 (2), veab064. doi:10.1093/ve/veab064

Oude Munnink, B. B., Nieuwenhuijse, D. F., Stein, M., O'Toole, A., Haverkate, M., and Mollers, M. (2020). Dutch-Covid-19 response: rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in The Netherlands. *Nat. Med.* 26 (9), 1405–1410. doi:10.1038/s41591-020-0997-y

Oude Munnink, B. B., Worp, N., Nieuwenhuijse, D. F., Sikkema, R. S., Haagmans, B., Fouchier, R. A. M., et al. (2021). The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. *Nat. Med.* 27 (9), 1518–1524. doi:10.1038/s41591-021-01472-w

Parkin, N. T., Avila-Rios, S., Bibby, D. F., Brumme, C. J., Eshleman, S. H., Harrigan, P. R., et al. (2020). Multi-laboratory comparison of next-generation to sanger-based sequencing for HIV-1 drug resistance genotyping. *Viruses* 12 (7), 694. doi:10.3390/v12070694

Quick, J. (2020). nCoV-2019 sequencing protocol v3 (LoCost). *protocols*. doi:10.17504/protocols.io.bp2l6n26rgqe/v3

Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., et al. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* 12 (6), 1261–1276. doi:10.1038/nprot.2017.066

Reusken, C., Broberg, E. K., Haagmans, B., Meijer, A., Corman, V. M., Papa, A., et al. (2020). Laboratory readiness and response for novel coronavirus (2019-nCoV) in expert laboratories in 30 EU/EEA countries, January 2020. *Euro Surveill.* 25 (6), 2000082. doi:10.2807/1560-7917.ES.2020.25.6.2000082

Rossen, J. W. A., Friedrich, A. W., Moran-Gilad, J., Genomic, E. S. G. f., and Molecular, D. (2018). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clin. Microbiol. Infect.* 24 (4), 355–360. doi:10.1016/j.cmi.2017.11.001

Shu, Y., and McCauley, J. (2017). GISAID: global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* 22 (13), 30494. doi:10.2807/1560-7917.ES.2017.22.13.30494

SoRelle, J. A., Wachsmann, M., and Cantarel, B. L. (2020). Assembling and validating bioinformatic pipelines for next-generation sequencing clinical assays. *Arch. Pathol. Lab. Med.* 144 (9), 1118–1130. doi:10.5858/arpa.2019-0476-RA

Tyson, J. R., James, P., Stoddart, D., Sparks, N., Wickenhagen, A., Hall, G., et al. (2020). Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv*, 2020.09.04.283077. doi:10.1101/2020.09.04.283077

Vogels, C. B. F., Breban, M. I., Ott, I. M., Alpert, T., Petrone, M. E., Watkins, A. E., et al. (2021). Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2. *PLoS Biol.* 19(5), e3001236. doi:10.1371/journal.pbio.3001236

Wegner, F., Roloff, T., Huber, M., Cordey, S., Ramette, A., and Egli, A. (2022). External quality assessment of SARS-CoV-2 sequencing: an ESGMD-SSM pilot trial across 15 European laboratories. *J. Clin. Microbiol.* 60 (1), e0169821. doi:10.1128/JCM.01698-21

TYPE Hypothesis and Theory
PUBLISHED 11 September 2024
DOI 10.3389/fmolb.2024.1348091

Check for updates

OPEN ACCESS

EDITED BY
Hem Chandra Jha,
Indian Institute of Technology Indore, India

REVIEWED BY
Ondrej Sobek,
Laboratory for CSF and Neuroimmunology, Topelex, Czechia
Pedro José Carlos Rondot Radío,
University of Buenos Aires, Argentina

*CORRESPONDENCE
Hansotto Reiber,
✉ csf.lcr@horeiber.de

RECEIVED 01 December 2023
ACCEPTED 24 July 2024
PUBLISHED 11 September 2024

CITATION
Reiber H (2024) Disease-related data patterns in cerebrospinal fluid diagnostics: medical quality versus analytical quantity.
Front. Mol. Biosci. 11:1348091.
doi: 10.3389/fmolb.2024.1348091

Disease-related data patterns in cerebrospinal fluid diagnostics: medical quality versus analytical quantity

Hansotto Reiber*

CSF and Complexity Studies Form, University Goettingen, Goettingen, Germany

Cerebrospinal fluid (CSF) diagnostics is characterized by the biologically relevant combination of analytes in order to obtain disease-related data patterns that enable medically relevant interpretations. The necessary change in knowledge bases such as barrier function as a diffusion/CSF flow model and immunological networks of B-cell clones and pleiotropic cytokines is considered. The biophysical and biological principles for data combination are demonstrated using examples from neuroimmunological and dementia diagnostics. In contrast to current developments in clinical chemistry and laboratory medicine, CSF diagnostics is moving away from mega-automated systems with a constantly growing number of individual analyses toward a CSF report that integrates all patient data. Medical training in data sample interpretation in the inter-laboratory test systems ("EQA schemes") has become increasingly important. However, the results for CSF diagnostics (EQAS from INSTAND) indicate a crucially misguided trend. The separate analysis of CSF and serum in different, non-matched assays and extreme batch variations systematically lead to misinterpretations, which are the responsibility of the test providers. The questionable role of expensive accreditation procedures and the associated false quality expectations are discussed. New concepts that reintegrate the medical expertise of the clinical chemist must be emphasized along with the positive side effect of reducing costs in the healthcare system.

KEYWORDS

external quality control schemes, cerebrospinal fluid diagnostics, disease-related data patterns, Reiber diagram, blood–brain barrier functions, immunological networks, biophysics in medicine, accreditation system

1 Introduction

The analysis of cerebrospinal fluid (CSF) for the diagnosis of neurological diseases has always been a particular challenge for clinical chemistry (Reiber, 2016c). The small extraction volume and the low analyte concentrations in the CSF required an improvement in the sensitivity of the analytical methods commonly used in clinical chemistry. The interpretation of the CSF data then became particularly challenging due to the need to differentiate between the fractions in the CSF originating from the blood and those originating from the brain, e.g., immunoglobulins. This led to the combined analysis of the CSF sample with the corresponding blood sample of the patient and the calculation of their ratio as the CSF/serum concentration quotient. This was the beginning of the evaluation concepts of combined data, a discussion that has now lasted 60 years. The linear index, the ratio of the serum protein quotient to the albumin quotient as a reference, remained popular despite the empirically and biophysically based non-linear relationships, e.g., those represented as hyperbolic lines in the quotient diagrams. The progress of neurochemical diagnostics (Wildemann et al., 2010; Reiber, 2016a; 2016b, 2016c) was initially less due to a growing number of new analytes than due to a biologically and medically relevant compilation of data patterns.

The additional introduction of a cumulative CSF data report, which integrated the patient clinical data in 1979 (Reiber, 2016c), became a model for other disciplines of clinical chemistry. A recent tutorial CSF App (Albaum and Reiber, 2024) illustrates this development of disease-related data patterns in CSF diagnostics (Wildemann et al., 2010; Reiber, 2016a; Reiber, 2016b).

The initial development of CSF diagnostics in the 1990s would not have been so successful without the support of Beckman and Dade Behring, the suppliers of the automated nephelometer machines. In the meantime, a decisive role change took place: industrial companies run the analytical invention according to their own rules, which are based more on their financial interests than on medical needs. This means disadvantages for the analytical quality of the combined CSF and serum analysis, as well as an explosion in analytical costs without any corresponding medical benefit. The certification and accreditation business with its high costs also contributes to the loss of quality in clinical neurochemistry as the providers of established online analysis software cannot afford to maintain their service.

1.1 Topics of the contribution

1. In CSF diagnostics, the following principles are used that can be generalized to improve the quality of laboratory medicine:
   - Quotient formation (CSF/serum proteins and dementia marker proteins).
   - Coefficients of variation (CVs) in CSF and blood (brain- and blood-derived proteins).
   - Immunoglobulin class patterns with brain-specific preconditions.
   - External quality analysis schemes (EQASs) with additional interpretation of medically relevant data patterns.

The current development in clinical neurochemistry is critically discussed with reference to the database of the INSTAND external quality assessment schemes (EQAS) for CSF (Reiber, 1995) and the current certification practice.

2. The integration of disease pathology and current knowledge bases must be part of any quality control of medical laboratory data. The second main aspect of this article is, therefore, the presentation of the medical and biological knowledge bases relevant for the plausibility control and, finally, the correct interpretation of diagnostic data. The obvious deficits in the practice of diagnostic interpretations (Uhr, 2024) and the subsequent lack of adequate therapeutic consequences make this an important goal (Reiber, 2024).

1.2 Current knowledge bases

The review of the following topics presents some essentials of clinical neurochemistry in particular and clinical chemistry in general, based on a recent publication (Reiber, 2024):

- The diffusion-flow model of blood–brain barriers
- The immunological networks (B-cell-based and pleiotropic cytokines).
- Biophysics and complexity approach in medical diagnostics.

In times of AI-based big data analysis, the call for a shift from a growing number of individual analytes that lack rational argumentation to a limited amount of functionally linked data seems to be swimming against the tide. However, it is not. In medicine, there are no sufficiently large datasets for disease group statistics with deep learning approaches; conversely, the perspective for AI with machine learning is to integrate an existing secure knowledge base to ensure the reality of categorizations (Wahlster, 2021).

With the integration of current knowledge bases, CSF diagnostics shows the relevance of data coupling to obtain a disease model-based selection of diagnostic datasets. The increasing acceptance of complex system approaches in diagnostics and EQASs in CSF show the foundations of a medical-based quality control that goes beyond the usual external accuracy control of individual analytes.

These aspects could contribute to the perspectives in medical laboratory diagnostics and restore the medical competence of clinical chemists.

2 Change of a paradigm: blood–brain and blood–CSF barrier functions

The barrier function is a fundamental topic of this work. A shift from mechanical, linear models to dynamic, non-linear biological functions requires more fundamental knowledge in medicine.

The most glaring example of a necessary paradigm change is the dysfunction of the blood–cerebrospinal fluid barrier, i.e., the pathological increase of serum protein concentrations in the cerebrospinal fluid. The serum protein concentrations in the CSF increase due to a reduced CSF flow rate, i.e., slowed removal, and not due to a hole in the barrier. However, the practice in scientific publications looks significantly different. A Google search under the keyword *blood–brain barrier* with 94,500 citations in the last 10 years is associated 78,900 times with “impairment,” 71,900 times with “breakdown,” and 56,400 times with “leakage” (Google search on 14 March 2023, period 2013–2023).

The idea of a barrier leak is as wrong as the expectation that a stone thrown into the lake will leave a hole in the water.

From an evolutionary point of view alone [see below and Reiber (2024)], such spontaneous instability of a structure that has existed in a variety of species for 500 million years is unlikely.

2.1 Barriers

The blood–brain and blood–cerebrospinal fluid barriers are morphologically different and also vary in different brain areas. The molecular passage between blood and extracellular fluid (ECF), the blood–brain barrier, and the passage from blood to CSF, the blood–cerebrospinal fluid barrier, are based on two basic transfer ways, the paracellular passage with facilitated or active transfer mechanisms and the intercellular passage for proteins, which depends on passive diffusion (Reiber, 2024). The biophysical principles in both barriers are the same, but the CSF has a 10-fold faster flow (turnover) than the ECF in the brain.

Barrier function for proteins: Serum proteins pass through the endothelial cell layer of the capillaries, which are reinforced with additional brain-specific structures, different in different areas of the brain. The molecular size-dependent restriction of the diffusion of proteins into the CSF is at a steady state with the elimination by CSF outflow (bulk flow). This steady state leads to very low–normal concentration ranges in the CSF, with approximately 1/100–1/3,000 (IgM) of serum concentrations for the most common proteins with the corresponding analytical problems.

2.2 Barrier dysfunctions

In many neurological diseases, the concentrations of serum proteins in CSF are pathologically increased, which is diagnostically characterized by the increased albumin CSF/serum concentration quotient, QAlb (Figure 1). The cause, a pathologically reduced CSF turnover, has three possible sources:

- The inflammation-related disturbance of CSF production in the ventricular plexus;
- A blockage in the subarachnoid space (tumor and stenosis); or
- A reduced outflow of CSF (swelling at the spinal roots in Guillain–Barré syndrome [GBS]).

The new paradigm for the barrier dysfunctions is the biophysically derived and patho-physiologically validated *diffusion-flow model* of the barriers (Reiber, 2003; Reiber, 2021a; Reiber, 2021b). It provides a rationale for the correct interpretation of a barrier dysfunction. Even the age-dependent increase in the normal QAlb value can be explained by reduced CSF production due to age-related changes in the choroid plexus. The increasing QAlb value, the barrier dysfunction, thus represents a reciprocal function of the pathologically decreasing CSF flow rate (Figure 1).

2.3 Discrepancies in interpretations

Figure 1 shows a consequence of the changing paradigms. The flow-related change in the local molecular flux, dependent on a non-linear concentration gradient across the barrier (Reiber, 2003; Reiber, 2021a), leads to non-linear, hyperbolic equations of the blood–CSF barrier function instead of linear correlations (Index, I, Figure 1). QAlb is the general reference value for the individual barrier function. It is a gift of nature that albumin, the largest serum protein fraction, is synthesized only in the liver and not also in the brain in pathological processes.

The discrimination between a protein fraction (e.g., IgG) diffusing from the normal blood across the barrier into the brain and a pathological protein fraction synthesized in the brain (intrathecal IgG fraction) is characterized by a hyperbolic limit function either in quotient diagrams (Qlim in Figure 1) or as a mathematical equation (Reiber, 2020). The still frequently assumed linear relationship (Index, I, in Figure 1) leads to interpretation errors. Calculation software for numerical and graphical statistical treatment of CSF data in diseases is available (free software from www.albaum.it), as explained by Reiber (2020).

The example of the blood–brain barrier function shows that without an understanding of the physics of diffusion and the stability of biological structures through material self-organization (see below), an adequate interpretation of the empirical data could be missed.

3 Neuroimmunology and immune networks

3.1 Immune reactions in the brain

On one hand, this example of a changed knowledge base replaces the clonal selection model of immunity an on the other hand, the idea that the brain is immunologically isolated. The immune system reacts as a network of B-cell clones, and the brain is linked to the immune and endocrine systems by pleiotropic cytokines. This leads to a new interpretation of the development of autoimmune diseases and immune system-associated pathologies such as bipolar spectrum disorders in psychiatry, which miss classical signs of immune reaction in the CSF.

Interpretations of the immune response in brain need to consider the following processes:

- All antibody-forming B cells in the brain migrate from the blood across the barrier into the brain and proliferate in the perivascular lymphocyte cuffs.
- The brain does not have an isotype switch (IgM to IgG class), which only happens in the lymph nodes.
- Antibody maturation (increase in avidity) happens only by selection of B-cell clones in the lymph nodes.
- The specificities of the polyspecific antibodies locally vary due to the random immigration of B cells of different specificities (the Ig class and measles, rubella, and varicella-zoster [MRZ] antibody patterns are different in the CSF and aqueous humor of the same individual).
- Immunocompetent cells of the CNS share cytokine receptors that link the nervous system to the immune and endocrine systems by pleiotropic cytokines.
- The innate immune system produces important components [e.g., of the complement system (Reiber et al., 2012)] in the brain, but their functions in disease development and defense (Jack et al., 2001) are not well understood.
- The polyspecific nature of any systemic and intrathecal immune response is the base for the diagnostics of chronic immune system-associated diseases (Reiber, 2024)

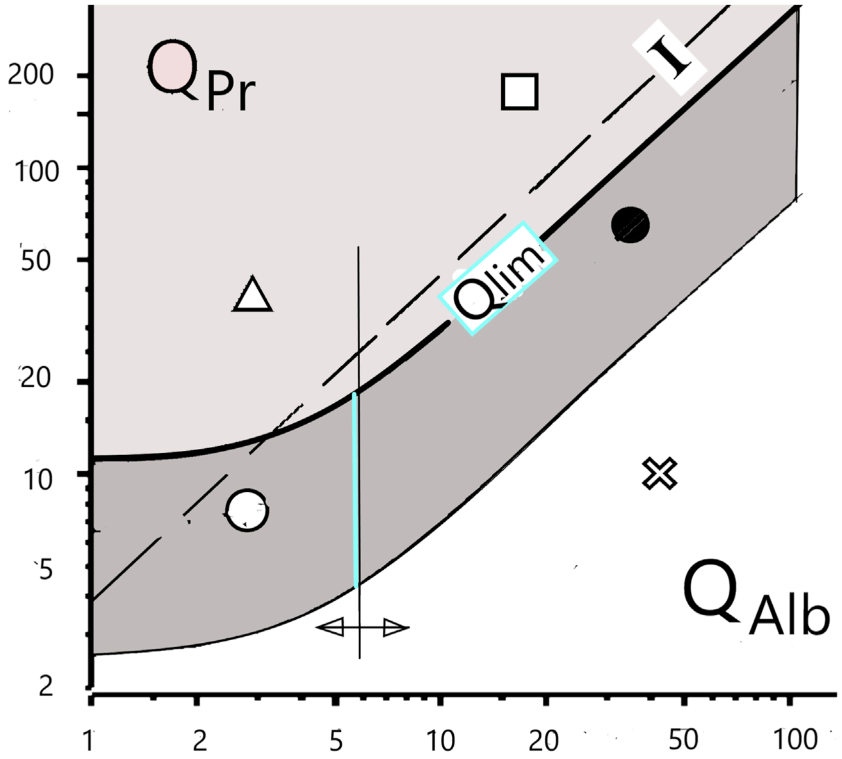


FIGURE 1
Differentiation of blood and brain fractions of proteins in the cerebrospinal fluid. The hyperbolic borderline (Qlim) is the result of the pathologically altered CSF flow rate, which has a non-linear influence on the local diffusion flow of proteins at the barrier interface to the CSF space. The previous assumption of a linear relationship ("Index I") is subject to errors of interpretation. The still frequent interpretation of a dysfunction of the blood–cerebrospinal fluid barrier, i.e., an increased QAlb, as a leakage in the barrier structure must be replaced by the new paradigm of a reduced cerebrospinal fluid turnover with an unchanged barrier structure.

3.2 Immunological networks

The idea of the immunological network dates back to 1975, with numerous theoretical approaches being developed in the 1980s (Varela and Coutinho, 1989; De Boer and Perelson, 1991). These theories extended the classical view that the adaptive immune system reacts to invading substances by producing antigen-specific antibodies (clonal selection) to rid the body of invaders. With the discovery of anti-idiotypic antibodies that recognize the body's own antibodies, lymphocyte-based networks became relevant. With the discovery of the polyspecific nature of the immune response in CSF, CSF research has made a major contribution to the practical significance of these theories. Felgenhauer et al. (1985) first described measles virus antibodies in the CSF of MS patients. Meanwhile, the observation that all immune responses, whether acute or chronic, involve polyspecific antibody synthesis (Reiber et al., 1998; Reiber, 2017a) has made the network property of the immune response an obligatory knowledge base for the interpretation of antibody data. In addition to this B-cell network, the trans-organ cytokine network associated with immunocompetent brain cells has also become important for the understanding and diagnosis of diseases related to the immune system, especially chronic diseases (Reiber, 2017a; Reiber, 2017b; Reiber, 2024).

3.2.1 Specific and polyspecific antibody synthesis in brain

The combined intrathecal MRZ antibody reaction in multiple sclerosis (Reiber et al., 1998), which is now of diagnostic relevance due to its strikingly high frequencies (Table 1), has also been reported in other chronic diseases (Hottenrott et al., 2015).

A fundamental new understanding came with the discovery in CSF diagnostics that in the case of an immune reaction to a specific antigen, as in herpes encephalitis (HSV) and subacute sclerosing panencephalitis SSPE (measles) (Jacobi et al., 2007), the disease-related specific antibodies account for only a small proportion of the total intrathecal antibody synthesis (Table 2). Nevertheless, the causative, specific antibodies have a 40–60-fold higher intrathecal synthesis rate than the polyspecific antibodies in chronic diseases (Jacobi et al., 2007).

In addition, the polyspecific antibodies differ from the specific antibodies by their generally higher avidity (Gharavi et al., 1996) (Table 2) as a consequence of the longer time interval for antibody maturation available in the case of chronic diseases.

3.2.2 Connectivity of B-cell clones in blood

The polyspecific immune reaction as a general feature of the immune response can also be detected in the blood. A representative example comes from a brilliant study of GBS patients (Terryberry

TABLE 1 Polyspecific antibody synthesis in the CNS in multiple sclerosis. Mean frequencies of intrathecally synthesized antibodies in multiple sclerosis patients against measles (M), rubella (R), varicella zoster (Z), herpes simplex (H), chlamydia (Chl), human herpes virus 6 (HHV6), toxoplasmosis (Tox), *Borrelia* (Bo), and double-stranded DNA (ds D). The mean values depend crucially on the respective MS cohort, depending on the intensity of intrathecal synthesis (Reiber et al., 1998). The patterns of polyspecific antibodies in the CSF vary from patient to patient.

Ab-species	M	R	Z	H	Chl	HHV6	Tox	Bo	ds D
Frequency in MS (%)	78	60	55	28	30	20	10	<25	19

TABLE 2 Comparison of intrathecal antibody responses to specific and polyspecific immune reactions. Calculated as specific intrathecal fraction Fs and expressed as the percentage of specific antibodies in relation to total intrathecally synthesized IgG (Jacobi et al., 2007). Measles-Ab and HSV-Ab are CSF values in subacute sclerosing panencephalitis, herpes simplex encephalitis, and multiple sclerosis. Rubella-Ab was determined in the aqueous humor (AH) of the eye in the Fuchs heterochromic cyclitis (FHC) and uveitis/periphlebitis of the eye as an MS symptom, MS(U).

Antigen	Specific Ab (%)		Polyspecific Ab (%)	
Measles (CSF)	18,8	SSPE	0,52	MS
Herpes s. (CSF)	8,9	HSVE	0,14	MS
Rubella (AH)	2,6	FHC	0,06	MS (U)

et al., 1995) with individually randomly elevated antibody and autoantibody titers in the blood. The pattern of elevated titers of polyspecific antibodies (Ab) is individually variable in an analysis set of 22 antibody assays in the 56 GBS patients (Terryberry et al. 1995), with between 1 and 13 simultaneously elevated titers, suggesting a Gaussian distribution for the arbitrary test set when the data are re-evaluated (Reiber, 2017a). The true network depth [connectivity (De Boer and Perelson, 1991)] in the individual patient, which includes autoantibodies, cannot be determined from the analyzed, arbitrary, and very limited test set.

The connectivity of B-cell clones was also demonstrated in a direct dynamics study.

The correlation of daily fluctuations in antibody concentrations in the blood of control subjects (Figure 2) also shows that connectivity is a feature of any normal immune response (Heitmann, 2002). Some Ab species fluctuate in a coupled manner, and others, independently. In both patients shown in Figure 2, the fluctuations in measles and rubella antibody concentrations are correlated with each other but not with dsDNA autoantibodies or mumps antibodies.

In addition to the specific antibodies, every acute infection with a specific antigen also leads to increased polyspecific antibody synthesis by a large number of other, already existing B-cell clones against other microorganisms, even autoantigens.

3.3 Connectivity-based interpretations—chronic immune reactions

Connectivity in the immune system explains many frequently misinterpreted observations. Only through the polyspecific activation of different coupled B-cell clones can lifelong immunity be maintained, as observed with measles immunization. The short lifespan of the B cells would not allow this if the specific B-cell clones were not maintained in the system through constant polyspecific activation.

The individual network depth of the connected B-cell clones determines the duration and efficiency of immunization (wild-type infection or vaccination). Connectivity can also be the cause of an autoimmune reaction by activating an existing B-cell clone for autoantibodies. When low levels of autoantibodies that provide immune tolerance are upregulated by a factor of 50, an autoimmune response appears as an undesirable damage associated with the individual immune response.

The Danish Disease Register (Nielsen et al., 2016) shows that autoimmune diseases and symptoms such as chronic fatigue syndrome/ME can occur a few weeks after the occurrence of infectious diseases. This also corresponds to post-Lyme or post-COVID symptoms. These correlations of non-specific, delayed immune reactions also make it clear that vaccinations can act as polyspecific triggers and, thus, also make the largely suppressed connection of the Gulf War illness with the 8-fold vaccinations in the first Gulf War plausible (Reiber and Davey, 1996; Rook and Zumla, 1997; Nielsen et al., 2016; Reiber, 2017b). Post-COVID syndrome has also been observed after vaccination, albeit less frequently than after wild-type infections. These associations offer a new approach to chronic diseases associated with the immune system that were previously difficult to diagnose (Reiber, 2024).

3.4 Cytokines, the trans-organ network

In the brain, cytokines are produced in immunocompetent cells, astrocytes, microglia, and endothelial cells. They can have both pro-inflammatory and anti-inflammatory effects. This is an important aspect of the immune response in the brain as a self-organizing local process, which is thought to be involved in various psychiatric chronic diseases (Bechter et al., 2010; Bechter, 2020; Runge et al., 2021).

The network is formed by the three properties of the different cytokines:

- Functional pleiotropism.
- Functional redundancy.
- Up and downregulation.

Pleiotropism refers to the binding of a cytokine to receptors in different organ systems. This creates a link between the three central control systems of the organism: nervous system, endocrine (hormonal) system, and immune system. Functional redundancy means that different cytokines can have the same effect in one and the same organ.

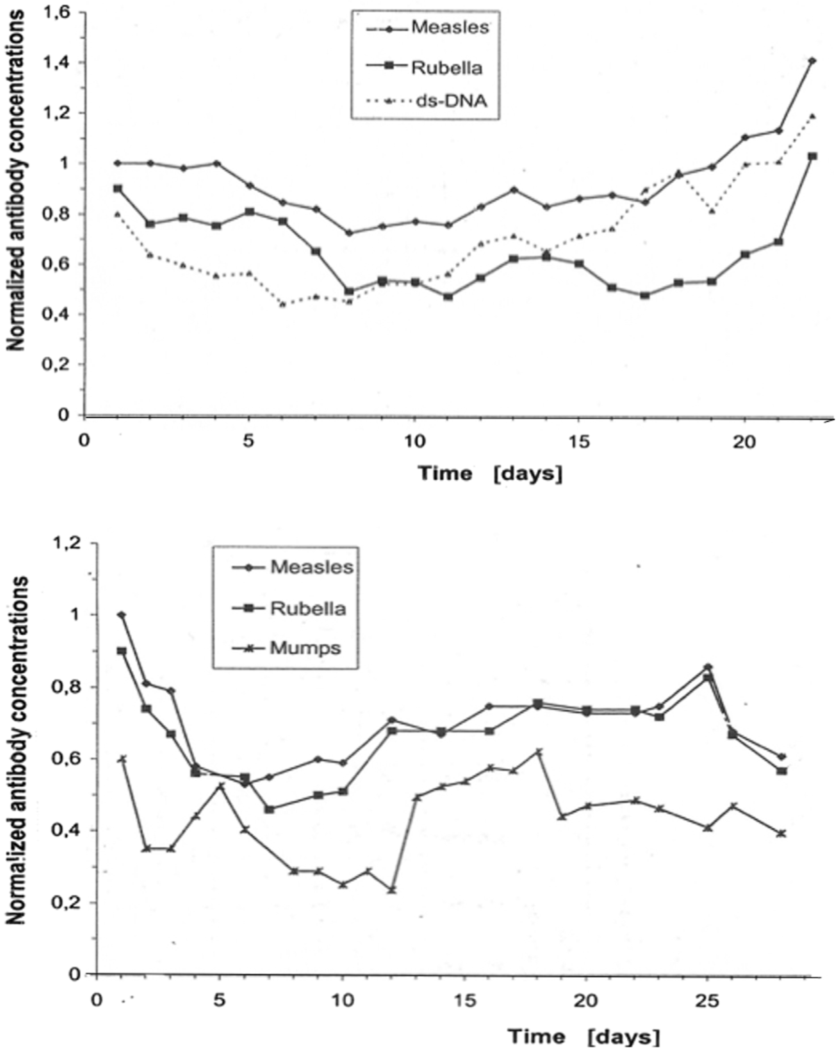


FIGURE 2
Antibody dynamics in the blood of patients without inflammation. The data originate from daily routine blood samples in the neurological intensive care unit of patients with a non-inflammatory disease (stroke) (Heitmann, 2002). These concentrations correspond to normal titers, i.e., they are below the threshold for nosocomial infections. The collected frozen samples were measured together in one analytical run (microtiter plates) to avoid bias due to inaccuracies in day to day imprecision (intra-assay imprecision CV = 3.5%). The concentrations of the individual antibody species were normalized to the first analytical value with C = 1. The total fluctuations in the blood proteins were controlled to exclude blood volume fluctuations as a cause. The curves are shifted slightly parallel along the *Y*-axis to better visualize the coupled daily concentration fluctuations.

Since one and the same cytokine can also be upregulating and downregulating depending on its concentration, a very complex network is created that is influenced by different body systems. This explains experiences that have led to the establishment of research concepts such as psychoneuroimmunology and diseases such as “stress-induced para-inflammation” (Bechter, 2020).

The involvement of cytokines in the initial processes of brain inflammation via the choroid plexus (Castellani et al., 2023) is also fundamental information for understanding the frequent barrier dysfunction in neurological diseases as a consequence of reduced cerebrospinal fluid production in the plexus.

The inclusion of cytokine analysis must become a fundamental requirement in neurology and psychiatry in order to explain pathologies associated with the immune system.

4 Biophysics in medicine

4.1 Complexity and Selforganization of Matter

The above examples of barrier function and the immunological network have highlighted a deficit in medical development. This becomes even clearer when we consider the unsuccessful therapeutic research into chronic diseases. There is no chronic disease for which we have a causal therapy, be it high blood pressure, type 2 diabetes, multiple sclerosis, autoimmune diseases, post-infectious chronic fatigue syndrome, heart failures, or the many types of cancer. This is primarily the result of the lack of complexity of the predominantly linear disease models, which are based on simple cause–effect relationships.

Fifty years ago, glycolysis oscillation (Hess and Boiteux, 1980; Gerok et al., 1989; Reiber, 2024) described how even in the smallest metabolic process of the cell, oscillating regulatory states can spontaneously change from one rhythm to another due to non-linear functions. This phenomenon is explained by complexity science as a change between attractors.

In addition to obtaining a better understanding of complex systems (Gerok et al., 1989; Goodwin, 2001; Reiber, 2007), we need to move from Jacob and Monod's refuted, 70-year-old model of the "genetic program" to the relevance of a phenotypic biology with a new view of epigenesis (Reiber, 2012). This also includes a fundamental recognition of the concepts of material self-organization in the context of non-equilibrium thermodynamics (Prigogine, 1997) or the non-linearity of biological processes in the context of complexity science (Mandelbrot, 1991). A short survey on these topics (Reiber, 2024) was published recently, providing more details.

There is no doubt that the improvement in medical diagnostics and therapy requires better biophysical disease models.

4.2 Dynamics of analyte concentrations

The daily fluctuations in the concentration of antibodies in the patient blood are shown in Figure 2. The coordinated variations indicate a common complex regulation. In patients with osteoporosis (Gerok et al., 1989), a disorder of calcium/parathyroid hormone regulation, the mean blood concentrations of calcium or parathyroid hormone remained unchanged compared to that in controls. Rather, the transitions changed from a normal, deterministic–chaotic time series of parathyroid hormone concentration in the blood to an almost constant concentration curve, indicating the loss of regulation. Analyzing a single blood sample may be useless for diagnosis, but a time series could provide it.

In contrast to analyzing individual samples, the electrocardiogram in heart disease provides a sufficiently large dataset to create time series (tachograms). The normally chaotic rhythm of the heartbeat (Reiber, 2024) changes under certain conditions to a time series, which can be analyzed for a lower complexity (change in fractal dimension) using complexity science methods (Mandelbrot, 1991; Reiber, 2024).

In general, time series (e.g., from electrocardiograms or electroencephalograms) are not available for clinical chemical diagnostics. However, these theoretical principles show that chronic diseases, in particular, can represent a stable state, i.e., have an attractor. This has consequences for diagnostics and therapy (Reiber, 2024).

This becomes clear, for example, in post-Lyme disease (Reiber et al., 2013): antibody findings must be interpreted differently, and treatment with antibiotics is pointless. The current discussion about post-COVID shares these problems.

CSF diagnostics cannot refer to the serial examination possible in blood, as a CSF puncture is rarely repeated unless for special reasons. This leads to a further concept for the integration of the complexity approach. The identification of disease-related data patterns can compensate, to a certain extent, for the complexity in the processes, as represented in the time series.

5 CSF diagnostics benefit from disease-related data patterns

Both the practical and theoretical principles of CSF diagnostics are documented in detail in various references (Reiber, 2016c; Lejon et al., 2003; Jacobi et al., 2007; Reiber et al., 2009; Wildemann et al., 2010; Reiber et al., 2013; Reiber, 2016a; Reiber, 2016b; Albaum and Reiber, 2024; Lewczuk et al., 2021; Reiber, 2024; DGLN 2024). Particular emphasis is placed here on the diagnostic principles that can be generalized for general quality improvement in laboratory medicine by identifying rational data combinations:

- Quotient formation and data variability (CV).
- Intrathecal immunoglobulin class patterns.
- Extended EQAS with the interpretation of data patterns.

5.1 Quotient formation—the role of biological variation

5.1.1 Definition and function

The CSF/serum concentration quotient (QAlb, QIgG, etc.) represents a normalized dimensionless CSF concentration that is independent of the individual variation in serum concentration [values between 0 and 1, which are used in practice due to the low CSF concentration in parts per thousand (× 10^{-3})].

This quotient is a biological relationship related to the laws of diffusion and not an arbitrary relationship such as in the calculation of protein concentrations from densitograms of serum electrophoresis. The CSF/serum quotient for proteins from the blood in the CSF excludes interpretation errors caused by greatly altered serum levels. For example, in patients with a serum IgM concentration increased up to 10-fold due to trypanosome infestation (Table 3) (Lejon et al., 2003), a subsequently increased CSF IgM concentration would be incorrectly interpreted as intrathecal IgM synthesis due to the high absolute values in the CSF. As the quotients are only determined by molecular size-dependent diffusion, the quotient remains normal despite extremely high serum concentrations (Table 3), provided that no barrier dysfunction or intrathecal synthesis contributes additional IgM.

The CSF/serum concentration quotient has the quality of a method-independent reference value, provided that the CSF and serum values are correctly analyzed in parallel (see below), i.e., the basic rule is no CSF analysis without serum analysis. The diagnostic uncertainty is hidden in the absolute values, not in the quotient calculated from biologically correlated data.

5.1.2 Coefficient of variation for the detection of data connections

For physically or biologically coupled values such as the CSF and serum concentrations of serum proteins (IgG, prothrombin, etc.; Table 4), the CV value of the quotient decreases compared to the absolute individual CSF CVs. The nominator of the quotient (CSF concentration) depends on the molecular size-dependent diffusion from a variable serum concentration (denominator of the quotient) and the individually varying length of the diffusion pathway and the CSF flow rate. Without the latter

TABLE 3 Blood concentrations and CSF/serum quotients in trypanosomiasis (stage 1 without brain involvement, parasitosis). Due to the constantly changing surface antigens of the trypanosomes, a new infection is always simulated with a corresponding new IgM class reaction. A 10-fold increased IgM concentration may contribute to the extremely low albumin protein production in the serum.

Proteins	Blood concentrations (g/L)		Mean quotients (x 10^3)	
	Normal control	Trypanos. 1. stage	Normal control	Trypanos. 1. stage
IgM	0.6–2.5	9–18	0,3	0,3
Albumin	35–55	23,2–33,6	5	5

TABLE 4 Inter-individual coefficients of variation, CVs, of CSF proteins from different sources. CSF, serum, and the CSF/serum quotients of each parameter are always from the same group with serum and CSF analyzed in the same analytical run. Comparison of blood-derived molecules [albumin, IgG, prothrombin (Reiber, 1995)], brain-derived neopterin (Kuehne et al., 2013), predominantly brain-derived proteins [transthyretin (Reiber, 2003)], and the leptomeningeal mannan-binding lectin (Reiber et al., 2012) or beta-trace protein (Reiber, 2003).

	Coefficients of variation, CV (%)						
	Alb	IgG	Proth	Neopt	MBL	TT	ßTrac
CSF	$29.3^{1)}$	46	33	10	66	6.4	24
Ser	8.9	35	21	25	146	18	28
Q x10^3	26.2	18	$22^{2)}$	21	81	23	25
N	53	23	18	26	13	28	28

[a] Limited age interval 50–59 years, mean 55 ± 3.2 years.

[b] Calculated from Qpr = 3 ± 0.7 at QAlb = 4.5 (Reiber, 1995).

parameters, we would have a universal constant that correlates with the diffusion coefficient D. The albumin quotient, which is freed from the variability of the serum concentration, still varies with the individually varying length of the diffusion pathway and the CSF flow rate. If, for example, we relate the IgG quotient to the QAlb, we remove these two fluctuations and obtain the universal non-linear hyperbolic relationship between QIgG and QAlb.

However, for independent parameters such as CSF and serum concentration of a brain-derived molecule (neopterin, MBL, and TT; Table 4), the CV of the quotient increases up to the sum of the two individual CVs. This contradicts the use of a quotient for data interpretation in the cases of brain-derived proteins.

5.1.3 Quotients in dementia diagnostics

The difference between biologically associated (A) and non-associated (B) analytes in process dynamics can be illustrated using the example of dementia diagnostics.

A. The beta-amyloid (1-42)/(1-40) ratio (=quotient).

The decrease in the beta-amyloid (1-42) concentration in the cerebrospinal fluid of a patient correlates with the formation of Alzheimer's plaques, with exceptions in the early phase. With the invention of the beta-amyloid (1-42)/(1-40) ratio, the inter-individual variations in total amyloid levels could be eliminated by referring to its main isoform, i.e., Aß40. This well-established ratio, which has recently been substantiated theoretically (Lewczuk et al., 2021), improved the diagnostic sensitivity compared to the determination of the beta-amyloid (1-42) concentration alone. This example is also mathematically a normalization that results in a relative change in beta-amyloid concentration corrected for individual differences in total amyloid levels. This has been empirically validated (Lewczuk et al., 2021) using patient groups and controls and shows that in the control group the 42/40 ratio has a 3-fold lower CV than the absolute values for Aß42. With a stronger decrease in the Aß42 concentration, the CV approaches both parameters [Aß42 ≈ Aß(42/40)], but it shows no inversion. In the critical range of the early Aß42 decrease, the ratio has the highest discrimination accuracy, i.e., sensitivity for the early detection of a pathological process.

B. Aß (1-42)/pTau protein ratio

Both proteins express Alzheimer's disease in different ways, but they provide different diagnostic information. Aß is reduced as an early expression of amyloidosis, the formation of amyloid oligomers, and the accumulation of amyloid-ß plaques. The Tau protein or pTau is increased later as an expression of the degenerative process on the neurons. This has two consequences: 1) in the same patient group, the CV of the ratio is greater than the individual CV of only Aß(1-42), i.e., the ratio is less sensitive, and 2) the ratio between the two analytes is variable depending on the course of the disease and is, therefore, susceptible to misinterpretation, e.g., if the individual patient happens to have a still normal but very low Tau or pTau protein concentration at the time of early Aß decrease. This would lead to a false negative interpretation (normal instead of elevated Aß42 levels). Despite the low sensitivity, however, this association can contribute to diagnostic specificity if the diagnostic question regarding the time of disease progression is framed correctly.

5.1.3.1 Summary

We described three versions of useful ratios between variables in the organism that improve analytical accuracy, diagnostic sensitivity, and diagnostic specificity:

- Biophysical coupling/normalization: decreased, constant CV.
- Biological coupling/normalization: decreased, not constant CV.
- Medical relevant pattern/association: increased CV

These differences need to be understood by regulatory bodies to finally end the useless discussions about the relevance of CSF/serum concentration ratios as part of guidelines as the ratio has less CV variation compared to absolute concentration values.

5.1.4 Research projects: source detection of proteins in CSF

Conversely, the data combination given in Table 4 could be used in research to characterize the origin of a new molecule in CSF based on its CV (Reiber, 2003; Kuehne et al., 2013). Further parameters for characterizing the origin of a molecule in CSF are the gradient depending on the molecular size, the CSF/serum concentration quotient, and the rostro-caudal concentration gradient, which increases in the case of blood-derived or leptomeningeal proteins (Reiber, 2003; Reiber, 2021b).

5.2 Immunoglobulin class patterns in neurological diseases

The disease-typical reaction patterns of intrathecal IgG, IgA, and IgM syntheses differ from the uniform immunoglobulin dynamics in the blood due to the lack of an isotype switch in the brain and the local character of the intrathecal immune response with a barrier passage of the immune cells originating from the blood. By calculating the relative intrathecal fractions of the total immunoglobulin in the CSF (intrathecal fractions, IF, shown as % lines of intrathecal Ig in Figure 3), disease-typical combinations of IgG, IgA, and IgM syntheses can be described in the quotient diagrams, which are also known as Reibergrams or Reiber diagrams (Figure 3).

5.2.1 Representative examples in quotient diagrams

The examples given in Figure 3 show neurotuberculosis, African trypanosomiasis, and two different courses of neurosyphilis:

1. Intrathecal IgA synthesis (Figure 3) in combination with increased CSF lactate, a barrier disorder, and a mean cell count are so highly specific that in European contexts, this is an important clue to an otherwise unexpectedly rare tuberculosis. This pattern eliminates the need to search for other pathogens, contributing to rapid diagnosis and treatment at a minimized cost.
2. The three-class reaction with dominant IgM response in trypanosomiasis (Lejon et al., 2003) (Figure 3) would also fit neuroborreliosis (Reiber et al 2013). However, confusion is impossible if the geographical differences are taken into account and the obligatory detection of trypanosomes in the blood with an extreme serum IgM concentration in trypanosomiasis is added (Table 3).
3. The different course of meningovascular and parenchymal neurosyphilis is shown in Figure 3 by an extremely strong additional intrathecal IgM synthesis (85% of the total IgM in the CSF). The intrathecal IgG synthesis with increased *Treponema* AI in neurosyphilis is found as a scar over decades despite adequate treatment and should not be interpreted as a sign of activity. Reactivation of neurosyphilis can only be recognized by the elevated serum IgM levelsynthesis as neurotuberculosis. Examples.

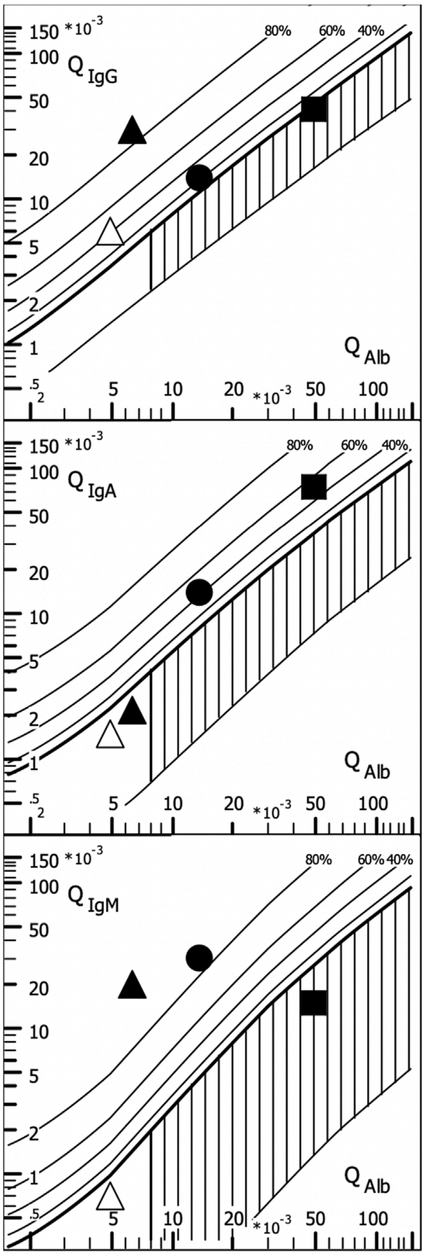


FIGURE 3
Quotient diagram for IgG, IgA, and IgM (Reibergram). Presentations of patient data for neurotuberculosis (squares), neurotrypanosomiasis (circles), and two patients with neurosyphilis with different courses (triangles). Neurotuberculosis shows a characteristic picture with isolated intrathecal IgA synthesis, whereas in neurosyphilis, conversely, a lack of IgA class reaction is more typical. Neurosyphilis shows a different pattern in the meningovascular course (open triangle) than in the parenchymatous course (filled triangle), with an additional strong IgM class reaction. In African trypanosomiasis, the dominant IgM class reaction with a frequent three-class reaction offers the highest specificity for brain involvement (sleeping sickness). For numerical characterization, the relative intrathecal fractions are used, such as IgG (IF) = [QIgG -Qlim (IgG)]/QIgG × 100 (%) (corresponding % lines in Figure 3).

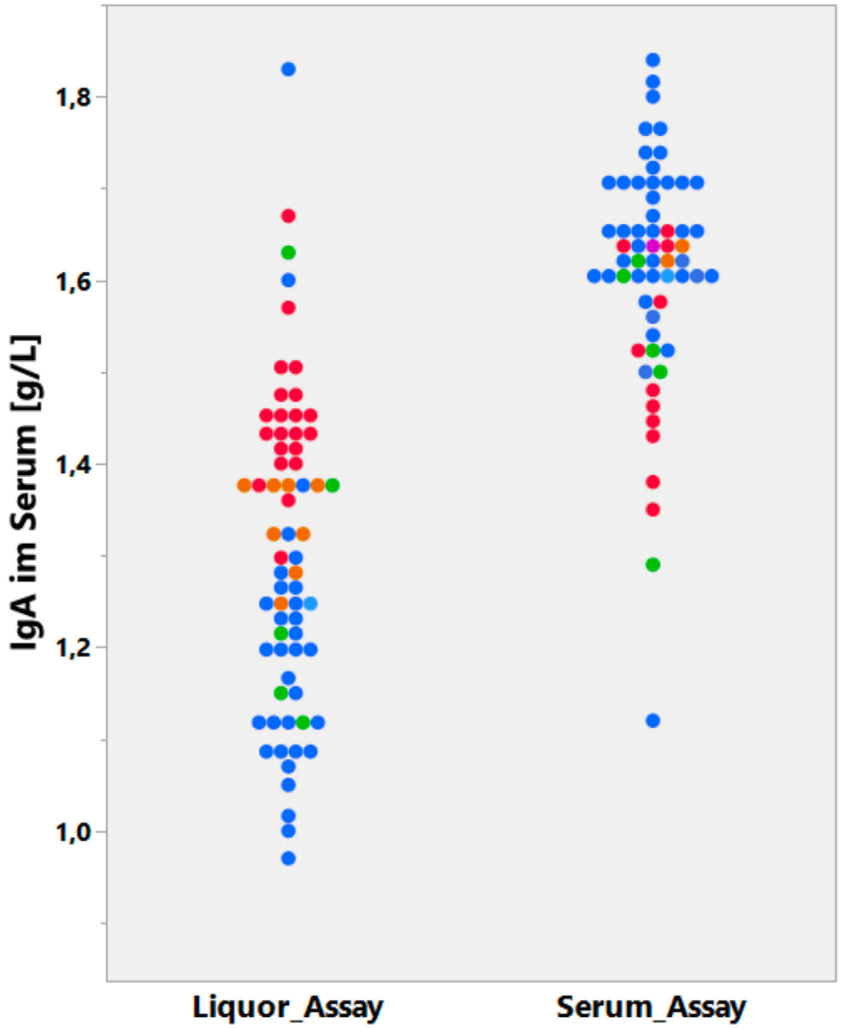


FIGURE 4
Data from the CSF EQAS of INSTAND, May 2022 (Uhr 2022). Method-dependent accuracy. Analysis of serum IgA in CSF or serum assay. Absolute IgA concentrations in serum samples are determined either in a serum assay (right column) or diluted in a CSF assay (left column). The CSF and serum assays that are from the same supplier (Siemens, blue) are not matched and must lead to a wrong CSF/serum IgA quotient.

Further examples are described with interpretations and comments in the CSF Tutor App (Albaum and Reiber, 2024), the reviews (Reiber, 2016c; Wildemann et al., 2010; Reiber, 2016a), and individual publications (Lejon et al., 2003; Reiber et al., 2009; Bechter et al., 2010; Kuehne et al., 2013; Reiber et al., 2013).

In general, this combined analysis of IgG, IgA, and IgM data can provide a number of important diagnostic clues, including the detection of analytical errors through plausibility checks (antigen excess in IgA analysis and blood contamination in CSF).

These associations of disease-related data also show that quality control via medical plausibility must be understood as an important complementary control to EQAS (see below).

5.2.2 Diagnostic sensitivity and specificity of quotient diagrams

The disease-specific Ig patterns have different sensitivities for different diseases and only gain their specificity with additional information. The disease-typical patterns refer to the first diagnostic puncture, which takes place at different times after the onset of the disease, depending on the disease, i.e., 1–2 days for bacterial meningitis, 1 week for viral encephalitis, or up to 3 weeks for neurotuberculosis, depending on the onset of symptoms.

The typical pattern and its sensitivity depend on the course of the immune response in the blood, the variable time for transmission of the microorganism through the barrier, and the locally variable onset of the disease in the brain. It is important to understand that an undetectable immune response in the CSF does not rule out disease as pathological processes far from the CSF space may be undetectable due to the long diffusion pathway to the CSF space. The carcinoembryonic antigen (CEA) is not detectable in the CSF if the CEA-producing tumor metastasis is localized in the frontal brain. This different localization of the pathological processes in the brain is one of the reasons that limits the detection sensitivity of the patterns.

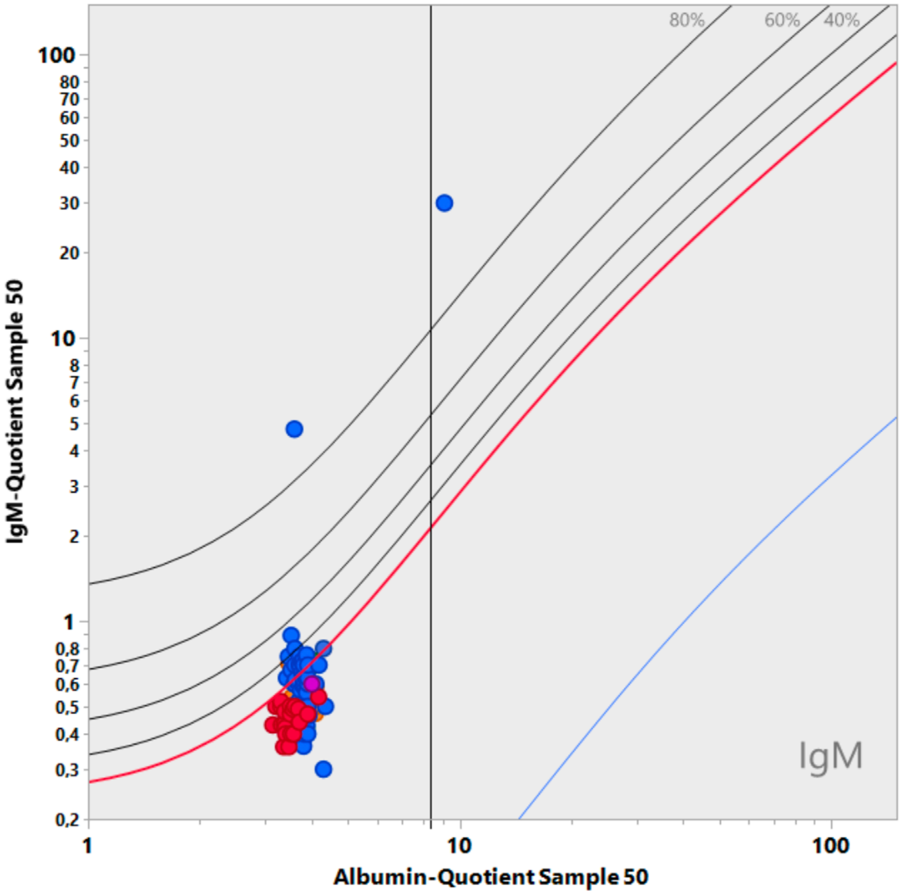


FIGURE 5
Data from the CSF EQAS of INSTAND, September 2022 (Uhr, 2022). Interpretation error in quotient diagrams. The absolute serum IgM values were analyzed in either the serum or CSF assay. If CSF and serum (diluted) were measured in one group in the same CSF assay, the CSF/serum IgM quotient led to a correct normal quotient, which is below Qlim in the diagram. The group that measured CSF in the CSF assay and serum in the serum assay yielded a wrong interpretation, indicating an intrathecal synthesis (QIgM greater than QIgMlim). Both assays were from the same supplier (Siemens).

The additional information in the quotient diagram pattern is an essential part of not only the specificity of this diagnostic approach, such as cell count, differential cell count, and lactate, but also the communication of the differential diagnostic questions. This integrating CSF data report [examples in the study by Reiber (2016c) and Reiber (2016a)], introduced in Göttingen in 1977 by the neurologists Sigrid Poser and Helmut Bauer, has become the standard for many laboratories.

However, the interpretation of the patterns undoubtedly requires training and knowledge to avoid diagnostic errors. A CSF Tutor app for the interpretation of disease-related patterns is available as free software (Albaum and Reiber, 2024). This also led to the integration of pattern interpretation into the external quality control systems for laboratory diagnostics as early as 1995 (Albaum and Reiber, 2024).

Disease-related data patterns should be considered as a perspective for analytical plausibility control in order to improve the specificity of diagnostics from a medical point of view, additionally increase the accuracy of interpretation, and, above all, reduce healthcare costs.

6 Quality control in the CSF laboratory

Details of quality control in the CSF laboratory have been reported by Wildemann et al. (2010) and in the collection of methods from the German CSF society (DGLN, 2024). These knowledge-based concepts were taken over only to a very restricted extent by the official quality control authorities (RiliBÄK). The improved EQAS of INSTAND (Albaum and Reiber, 2024) was a normative example, which is still the leading, albeit most demanding concept.

6.1 Medical data as part of quality control

The evaluation of combined data patterns as part of an EQAS with INSTAND had been introduced in the 1990s (Albaum and Reiber, 2024). In many surveys, we demonstrated that the quality of medical diagnostics does not depend on the smallest possible coefficient of variation of a method but on the interpretability of laboratory data by the clinical chemist

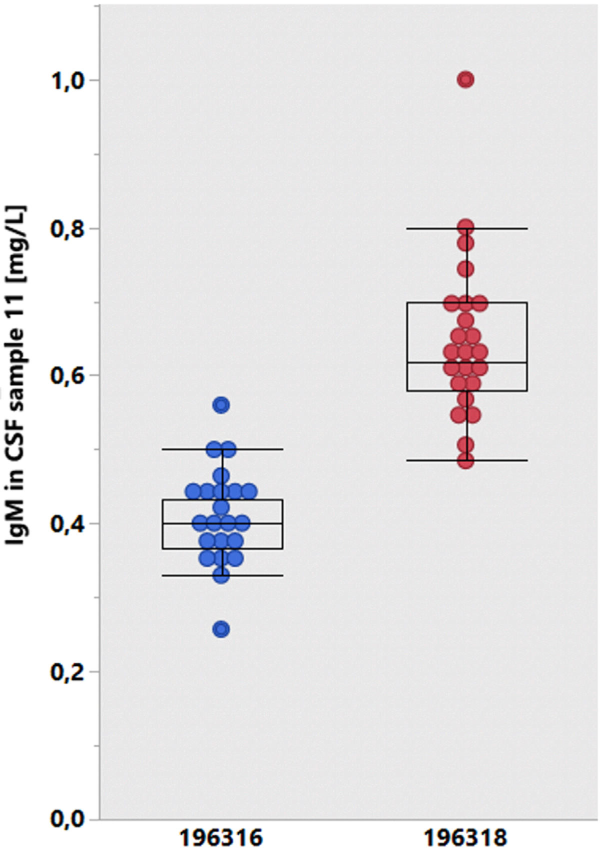


FIGURE 6
Lot-dependent test results of the INSTAND survey (Uhr, 2022, January 2020). The results were cumulated from participant groups according to the lot numbers of the Siemens Healthcare tests for IgM in CSF. A mean lot-to-lot difference of 50% is hardly reliable.

and, finally, the physician. The Göttingen neurologist Klaus Felgenhauer, who emphasized the importance of laboratory analyses, said, "If a laboratory finding does not fit into my clinical diagnostics, I ignore it." This can save lives, but it can also cause an alternative diagnosis to be missed. This aphorism became reality when a laboratory report of an isolated intrathecal IgM synthesis with a normal cell count and other normal protein concentrations was ignored as the clinicians had discarded an inflammatory process. So, the alternative interpretation as a non-Hodgkin lymphoma in the CNS was missed by the neurologists, and it needed a particular case conference to find the diagnosis with delay. To query these options of intrathecal IgM synthesis in an external quality control system contributes more to the quality of laboratory diagnostics than the correctness of absolute IgM concentration values. Some control institutions (Controllab, Brazil) cover this aspect of knowledge-based interpretations by an occasional question/answer test. INSTAND is hosting an online conference of the surveyor with the customers, in addition to the explicit medically oriented commentary, with the results reported to the participants.

6.2 Misleading industrial trends in clinical chemistry

6.2.1 CSF protein analysis

The ideal method is to measure the CSF and correspondingly diluted serum in the same assay with a calibration curve that is true to dilution (same recovery in all concentration ranges). Thankfully, this concept was originally developed by Beckman and Dade Behring for their nephelometer machines in cooperation with the Neurochemical Laboratory in Göttingen (head: H. Reiber). The immunochemical analysis with antigen-coated microtiter plates served as the reference method, which is by far the most sensitive method. In the meantime, other suppliers of test systems for basic CSF protein analysis participate in the survey (Siemens, Roche (turbidimetry), Beckman, Abbott, Binding Site, and others).

6.2.2 Actual analytical quality problems

The current, very serious analytical problems are demonstrated in several latest surveys (Uhr, 2022), with interpretations and comments by Manfred Uhr, the current surveyor of the CSF inter-laboratory test at INSTAND (Uhr, 2022). The following results focus on the performance of participants using Siemens analyzers, which form the largest group and also have the most problematic performance due to the conceptual decisions of the assay provider. There are three main problems.

1. Separate analysis of the CSF and serum samples of a patient in different assays.
2. Non-matching calibration curves in the CSF and serum assays (bad recovery).
3. Large variations from batch to batch in the CSF assay.

6.2.2.1 Unreliable recovery in unmatched CSF and serum assays

The problems start at the lowest level: assay manufacturers avoid the cost of matching the calibration of their CSF assay with the calibration of the serum assay to ensure the accuracy of the CSF/serum quotient (e.g., CSF/serum concentration quotient, QIgG). This is shown in Figure 4.

Figure 4 summarizes the serum IgA concentration measured by one group of participants with the serum IgA assay and (the same serum sample diluted) the CSF assay by another group (Uhr, 2022). The strongest difference in the recovery of the serum IgA concentration between CSF and serum assays was observed in participants using the Siemens assays (blue).

6.2.2.2 Misinterpretation due to unmatched CSF and serum assays

Figure 5 shows the fatal consequences of IgM serum analysis with unmatched assays for a CSF/serum sample pair. Participants with the Siemens assay who measured the serum with the serum assay received a quotient that was too high, which indicated a pathological IgM value (QIgM > QIgMLim). As IgA and IgG levels were normal in the patient, the result is misinterpreted

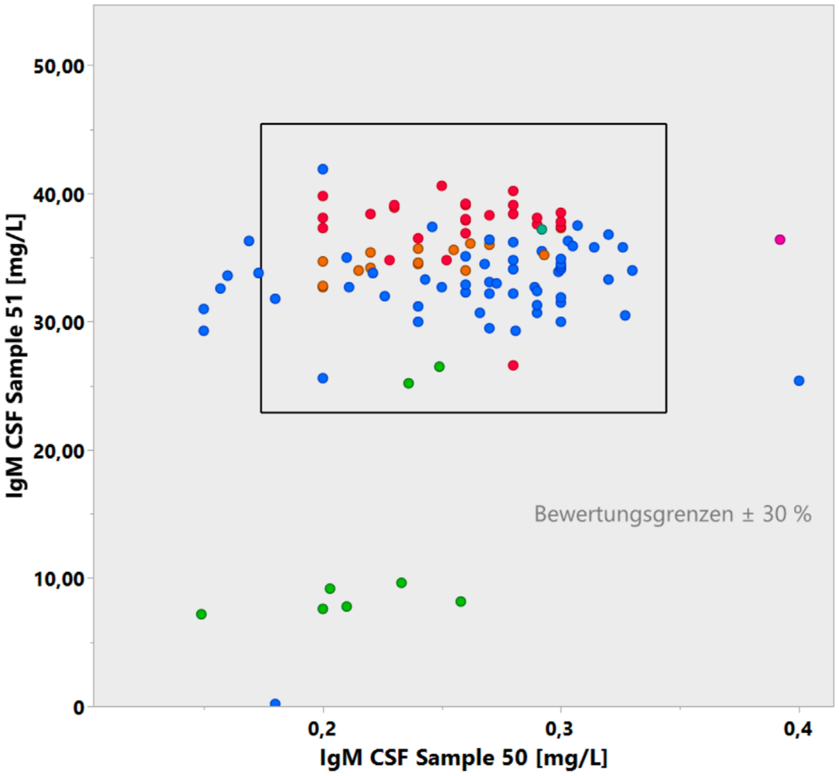


FIGURE 7
Youden plot of IgM concentrations in two CSF samples in the CSF EQAS (Uhr, 2022, September 2022). The accuracy of the lower concentrations in sample 50 with a larger coefficient of variation does not match the accuracy of sample 51.

as an inflammatory process or can even be interpreted as an indication of intrathecal lymphoma due to the isolated intrathecal IgM synthesis. In daily practice, this would lead to a serious health risk for patients and an extreme increase in costs in the clinic (see example above).

6.2.2.3 Inaccuracy from batch to batch

The second bias by the test provider results from the inadequate difference between batches of CSF tests (Figure 6). When analyzing the CSF samples with different batches from the provider (Siemens Healthineers), a difference of 50% was found between the two groups (Figure 6), which represents a problem of quotient accuracy, especially when the CSF and serum analyses are performed with different assays. Siemens made a conscious decision by canceling the certificate of the CSF assays for serum analysis. This manufacturer-dependent performance can lead to fatal consequences for the patient, as the interpretation example in Figure 5 shows.

Given the bias in analyzing CSF and serum samples in different assays (Figures 4, 5) and the amplifying bias due to different accuracies in different CSF assay batches (Figure 6), it would be reasonable to insist in principle that CSF and serum samples for protein analysis are measured in the same assay with reference to the same calibration curve and that the dilution accuracy of the common calibration curve is guaranteed, which is a basic problem, as shown in Figure 7.

6.3 Knowledge-based pattern interpretation as part of EQA systems

6.3.1 Structure of the survey

The example given in Figure 8 summarizes the three basic aspects of a modern CSF EQAS:

- Analysis of CSF and serum samples in a reliably matched assay procedure (see above).
- Knowledge-based interpretation of the combined data of medically relevant examples.
- CSF and serum samples provided for the survey must enable relevant pattern formation.

With the high IgG, IgA, and IgM concentrations in the CSF, in Figure 8, we obtain a relatively good agreement of the results from all survey participants. However, the interpretation of the IgA quotient requires knowledge-based awareness. Most of the QIgA results given in Figure 8 are below the discrimination limit, QIgA < QIgA(lim), and they were, therefore, interpreted together with the other quotients as barrier dysfunction without inflammatory process. Only 33% of participants recognized that QIgA > QIgG, implying intrathecal IgA synthesis, despite QIgA < QIgA(lim).

The knowledge-based explanation is as follows: the IgA quotient of 64.6 × 10^{-3} was greater than the IgG quotient of 51.6 × 10^{-3}. IgA has a larger hydrodynamic radius than IgG and, therefore,

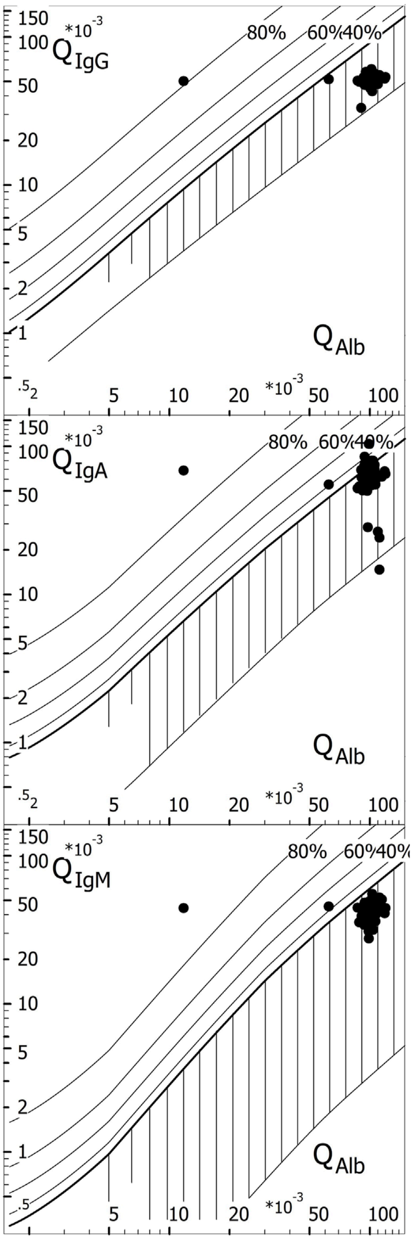


FIGURE 8
Knowledge-based interpretation of borderline intrathecal IgA synthesis as neurotuberculosis. Example in the EQAS (Uhr, 2022, survey 2021, October). The IgA quotient of 64.6 × 10^{-3} was greater than the IgG quotient of 51.6 × 10^{-3} (confirmed by 124/125 participants). Only 33% of all participants with a complete analysis (42/125) reported intrathecal IgA synthesis with interpretation as an inflammatory process. N = 12 participants calculated QIgA > QIgA (lim). Of the 113 participants with values < Qlim, only 30 (27%) recognized that if QIgA > QIgG, the interpretation must be intrathecal IgA synthesis. Thus, 67% of all participants who analyzed a complete Reibergram did not recognize an inflammatory process and, more crucially, would not have alerted the clinician to possible neurotuberculosis (Albaum and CSF APP). Even worse, 83% of all laboratories (243) that participated in the CSF survey, eventually limited to analyzing the total protein in the CSF, failed to alert the clinician.

must have the smaller quotient due to diffusion (flatter gradient along the diffusion barrier). The normal quotient sequence is QALB > QIgG > QIgA > QIgM. This molecular size-dependent interpretation of the quotients also applies in the case of possible blood contamination, be it artefactual due to puncture or cerebral hemorrhage.

Thus, 67% of the participants did not recognize an inflammatory process and, more importantly, would not have alerted the clinician to a possible case of neurotuberculosis if this was associated with an elevated CSF lactate level (Albaum and Reiber, 2024).

6.3.1.1 The corresponding practical case

We recently learned about the potentially fatal consequences for the patient from a practical case of an IgA analysis in which the very high CSF IgA concentration was beyond the analytical range. The machine in a central laboratory of a university hospital reported the value above the detection limit. The clinical chemist saw no need to measure again with a diluted CSF sample. This would have put her/him in the dilemma of using an assay with a procedure that deviated from the certified protocol. For the patient, this meant a delay in the diagnosis of neurotuberculosis (Figure 3), with 3 wasted weeks of precious treatment.

6.3.2 Samples for CSF testing in the EQAS

The availability of paired CSF and serum samples poses a particular problem in CSF testing.

In practice, residual CSF and serum samples are stored at 4°C after routine analysis and then pooled with different concentration ranges to enable different sample combinations. CSF was rarely obtained from a catheter that had a sufficient volume for all participants in the EQA scheme.

It should be noted that only by using real CSF samples that have an appropriate protein pattern (different protein ratios in CSF and serum due to diffusion through the barrier depending on molecule size) can such a pattern as shown in Figure 8 be created. The use of diluted serum instead of CSF (as practiced by some EQA schemes) does not allow the creation of disease-typical patterns (Albaum and Reiber, 2024). The stereotypical pattern with diluted serum completely deprives the examination of disease-related interpretation training and control.

7 Online interpretation software and the certification trap

7.1 Online evaluation of protein analysis

The concept of an integrating laboratory report, which presents all laboratory data of a patient together with the interpretation in a quotient diagram (Reiber, 2016c; Wildemann et al., 2010; Reiber, 2016a; Albaum and Reiber, 2024), made it seem obvious to couple the absolute values measured on the nephelometer machine with software that integrates the data directly into an online laboratory data report. These software developments by Albaum (2024) for Dade Behring (now Siemens Healthcare with another software manufacturer) and by Wormek (2024) for Beckman Nephelometers contributed significantly to the acceptance of the Reibergrams. Siemens offers the “Protis” program (Siemens, 2024) as part of an advanced diagnostic concept. However, the expensive accreditation of this long-established software program currently prevents providers from offering their software, e.g., for stand-alone solutions in individual laboratories, possibly in an international context.

7.2 Certification and accreditation

Test certification, originally introduced as laboratory marketing, has developed over the decades into a profit-orientated industry in its own right. This has had devastating consequences, particularly for the isolated solutions developed and offered for sale by smaller IT companies. Waiting several years for certification, e.g., by the TÜV (Official technical surveyance of cars in Germany), and, above all, the annual costs of 30,000–50,000 Euros are a deterrent. The additional personnel costs of the software manufacturer for the administrative requirements make, for example, the certification of a simple CSF software program that has been functioning for 30 years prohibitively expensive as these costs cannot be passed on to the user.

The fact that a specialist in metallurgy is sent to the institute for quality assurance in medical diagnostics (INSTAND eV) for accreditation is one of the symptomatic problems of this type of pseudo-inspection. This does not contribute to diagnostic accuracy. Instead, it results in the explosion of costs in the healthcare system. Ultimately, it also means that development in small IT companies is no longer financially viable, and the programs are therefore not even offered. Therefore, this certification practice is an obstruction to quality in medicine.

8 Perspectives

The combined data evaluation is a qualified quality control of the laboratory analysis. Qualified, i.e., disease-typical data combinations avoid arbitrary analytical search processes and significantly reduce analysis costs. A purely analytical accuracy testing, as is considered sufficient in most EQA programs, therefore, does not show the necessary qualities of a good laboratory in data processing and interpretation. The misleading trends of assay suppliers must be controlled by medical expertise. Clinical chemists and laboratory physicians are therefore called upon to regain the medical expertise that has been lost through industrial developments. This is the only way to achieve a change in direction from mass analysis to patient-orientated and, at the same time, more cost-effective diagnostics.

The knowledge and responsibility of the clinical chemist cannot be replaced or hindered by certification or accreditation procedures without risk to patients.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found here: reports to the CSF surveys, www.INSTAND-eV.de.

Ethics statement

Ethical approval was not required for the studies involving humans because the EQAS uses blood and CSF samples that are anonymized by pooling and need no reference to the diagnostics of individuals. The samples were residuals after diagnostic procedures. The studies were conducted in accordance with the local legislation and institutional requirements. The human samples used in this study were acquired as a by-product of routine care or industry. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

HR: writing–original draft.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Publication fees were paid by INSTAND e.V., Duesseldorf, Germany.

Acknowledgments

The author deeply appreciates the support by Prof. Dr. Manfred Uhr, München, and Prof. Dr. Piotr Lewczuk, Erlangen. Their competent comments helped improve this paper. The translation/language program deepl.com was used to control the writing performance.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Albaum, W. (2024). CSF-Lab. Labororganisation fur das Liquorlabor. Available at: http://www.albaum.it.

Albaum, W., Reiber, H., and CSF App (2024). *Free download from* www.album.it *(android, windows) or Apple store (iPhone).*

Bechter, K. (2020). The challenge of assessing mild Neuroinflammation in severe mental disorders. *Front. Psychiatry* 11, 773. doi:10.3389/fpsyt.2020.00773

Bechter, K., Reiber, H., Herzog, S., Fuchs, D., Tumani, H., and Maxeiner, H. G. (2010). Cerebrospinal fluid analysis in affective and schizophrenic spectrum disorders. Identification of subgroups with immune responses and blood-CSF barrier dysfunction. *J. Psychiatric Res.* 44, 321–330. doi:10.1016/j.jpsychires.2009.08.008

Castellani, G., Croese, T., Peralta Ramos, J. M., and Schwartz, M. (2023). Transforming the understanding of brain immunity. *Science* 380 (6640), eabo7649. Epub 2023 Apr 7. PMID: 37023203. doi:10.1126/science.abo7649

De Boer, R. J., and Perelson, A. S. (1991). Size and connectivity as emergent properties of a developing immune network. *J. Theor. Biol.* 149 (3), 381–424. doi:10.1016/s0022-5193(05)80313-3

DGLN (2024). Deutsche Gesellschaft für Liquordiagnostik und klinische Neurochemie Methodenkatalog. Available at: www.dgln.de.

Felgenhauer, K., Schädlich, H. J., Nekic, M., and Ackermann, R. (1985). Cerebrospinal fluid virus antibodies. A diagnostic indicator for multiple sclerosis? *J. Neurol. Sci.* 71, 291–299. doi:10.1016/0022-510x(85)90067-x

Gerok, W., Gerok, W., and Hirzel, S. (1989). "Ordnung und Chaos als Elemente von Gesundheit und Krankheit," in *Ordnung und Chaos in der unbelebten und belebten Natur* (Stuttgart: Wissenschaftliche Verlagsgesellschaft mbH).

Gharavi, A. E., and Reiber, H. (1996). in *Affinity and avidity of autoantibodies*. Editors A. J. B. Peter, and Y. Schoenfeld (Germany: Elsevier Science), 13–23.

Goodwin, B. (2001). *How the leopard changed its spots. The evolution of complexity*. Princeton, NJ: Princeton University Press.

Heitmann, S. (2002). *Dynamik der polyspezifischen Immunreaktion bei nosokomialen Infektionen [Doctoral thesis]*. Goettingen, Germany: University Goettingen.

Hess, B., and Boiteux, A. (1980). Spatial dissipative structures in yeast extracts. *Ber. Bunsenges. Phys. Chem.* 84, 392–398. doi:10.1002/bbpc.19800840420

Hottenrott, T., Dersch, R., Berger, B., Rauer, S., Eckenweiler, M., Huzly, D., et al. (2015). The intrathecal, polyspecific antiviral immune response in neurosarcoidosis, acute disseminated encephalomyelitis and autoimmune encephalitis compared to multiple sclerosis in a tertiary hospital cohort. *Fluids Barriers CNS* 12 (1), 27. doi:10.1186/s12987-015-0024-8

Jack, D. L., Klein, N. J., and Turner, M. W. (2001). Mannose-binding lectin: targeting the microbial world for complement attack and opsonophagocytosis. *Immunol. Rev.* 180, 86–99. doi:10.1034/j.1600-065x.2001.1800108.x

Jacobi, C., Lange, P., and Reiber, H. (2007). Quantitation of intrathecal antibodies in cerebrospinal fluid of subacute sclerosing panencephalitis, herpes simplex encephalitis and Multiple sclerosis: discrimination between microorganism-driven and polyspecific immune response. *J. Neuroimmunol.* 187, 139–146. doi:10.1016/j.jneuroim.2007.04.002

Kuehne, L. K., Reiber, H., Bechter, K., Hagberg, L., and Fuchs, D. (2013). Cerebrospinal fluid neopterin is brain-derived and not associated with blood-CSF barrier dysfunction in non-inflammatory affective and schizophrenic spectrum disorders. *J. Psych.Res.* 47, 1417–1422. doi:10.1016/j.jpsychires.2013.05.027

Lejon, V., Reiber, H., Legros, D., Djé, N., Magnus, E., Wouters, I., et al. (2003). Intrathecal immune response pattern for improved diagnosis of central nervous system involvement in trypanosomiasis. *J. Infect. Dis. (JID).* 187 (187), 1475–1483. doi:10.1086/374645

Lewczuk, P., Wiltfang, J., Kornhuber, J., and Verhasselt, A. (2021). Distributions of Aβ42 and aβ42/40 in the cerebrospinal fluid in view of the probability theory. *Diagnostics* 11, 2372. doi:10.3390/diagnostics11122372

Mandelbrot, B. (1991). *Die fraktale Geometrie der Natur*. Basel: Birkhauser.

Nielsen, P. R., Kragstrup, T. W., Deleuran, B. W., and Benros, M. E. (2016). Infections as risk factor for autoimmune diseases - a nationwide study. *J. Autoimmun.* 74, 176–181. doi:10.1016/j.jaut.2016.05.013

Prigogine, I. (1997) "The end of certainty," in *Time, chaos and the new laws of nature*. New York: The Free Press.

Siemens (2024). Automated CSF assessment [access: 21/01/2020]. Available at: http://www.siemens-healthineers.com/diagnostics-it.

Reiber, H. (1995). External quality assessment in clinical neurochemistry: survey of analysis for cerebrospinal fluid (CSF) proteins based on CSF/serum quotients. *Clin. Chem.* 41, 256–263. doi:10.1093/clinchem/41.2.256

Reiber, H. (2003). Proteins in cerebrospinal fluid and blood: barriers, CSF flow rate and source-related dynamics. *Restor. Neurology Neurosci.* 21, 79–96.

Reiber, H. (2007). "Die Komplexität biologischer Gestalt als zeitunabhängiges Konstrukt im Zustands-Raum. Zum naturwissenschaftlichen Umgang mit Qualitäten," in *D. Zeilinger (Hg): VorSchein, Jahrbuch der Ernst-Bloch- Assoziation* (Nürnberg: Antogo Verl), 39–61.

Reiber, H. (2012). Epigenesis and epigenetics-understanding chronic diseases as a selforganizing stable phenotype. *Psych. Brain Res.* 18, 79–81. doi:10.1016/j.npbr.2012.02.001

Reiber, H. (2016a). Cerebrospinal fluid data compilation and knowledge-based interpretation of bacterial, viral, parasitic, oncological, chronic inflammatory and demyelinating diseases: diagnostic patterns not to be missed in Neurology and Psychiatry. *Arq. Neuropsiquiatr.* 74, 337–350. doi:10.1590/0004-282X20160044

Reiber, H. (2016b). Knowledge-base for interpretation of cerebrospinal fluid data patterns - essentials in neurology and psychiatry. *Arq. Neuropsiquiatr.* 74, 501–512. doi:10.1590/0004-282X20160066

Reiber, H. (2016c). Cerebrospinal fluid diagnostics in Germany since 1950: developments in the GDR and FRG in the context of society and science. *Nervenarzt* 87, 1261–1270. doi:10.1007/s00115-016-0241-7

Reiber, H. (2017a). Polyspecific antibodies without persisting antigen in multiple sclerosis, neurolupus and Guillain-Barré syndrome: immune network connectivity in chronic diseases. *Arq. Neuropsiquiatr.* 75 (8), 580–588. doi:10.1590/0004-282X20170081

Reiber, H. (2017b). Chronic diseases with delayed onset after vaccinations and infections: a complex systems approach to pathology and therapy. *J. Arch. Mil. Med.* 5 (3), e12285. doi:10.5812/jamm.12285

Reiber, H. (2020). Software for cerebrospinal fluid diagnostics and statistics. *Rev. cubana Investig. Biomed.* 39 (3), 740.

Reiber, H. (2021a). Blood-CSF barrier dysfunction means reduced cerebrospinal fluid flow not barrier leakage: conclusions from CSF protein data. *Arq. Neuropsiquiatr.* 79, 56–67. doi:10.1590/0004-282X-anp-2020-0094

Reiber, H. (2021b). Non-linear ventriculo - lumbar protein gradients validate the diffusion-flow model for the blood-CSF barrier. *Clin. Chim. Acta* 513, 64–67. doi:10.1016/j.cca.2020.12.002

Reiber, H. (2024). Cerebrospinal fluid diagnostics in Neurology. Paradigm change in Brain barriers, Immune System and Chronic Diseases. Berlin: Springer.

Reiber, H., and Davey, B. (1996). Desert-storm-syndrome and immunization. *Arch. Intern. Med.* 156, 217. doi:10.1001/archinte.156.2.217

Reiber, H., Padilla-Docal, B., Jensenius, J., and Dorta-Contreras, A. J. (2012). Mannan-binding lectin in cerebrospinal fluid: a leptomeningeal protein. *Fluids Barriers CNS* 9, 17–24. doi:10.1186/2045-8118-9-17

Reiber, H., Ressel, C., and Spreer, A. (2013). Diagnosis of Neuroborreliosis-Improved knowledge base for qualified antibody analysis and cerebrospinal fluid data pattern related interpretations. *Neurol. Psychiatry Brain Res.* 19, 159–169. doi:10.1016/j.npbr.2013.10.004

Reiber, H., St, U., and Chr, J. (1998). The intrathecal, polyspecific and oligoclonal immune response in multiple sclerosis. *Mult. Scler.* 4, 111–117. doi:10.1177/135245859800400304

Reiber, H., Teut, M., Pohl, D., Rostasy, K. M., and Hanefeld, F. (2009). Paediatric and adult multiple sclerosis: age-related differences and time course of the neuroimmunological response in cerebrospinal fluid. *Mult. Scler.* 15, 1466–1480. doi:10.1177/1352458509348418

Rook, G. A., and Zumla, A. (1997). Gulf War syndrome: is it due to a systemic shift in cytokine balance towards a Th2 profile? *Lancet* 349 (9068), 1831–1833. [PubMed: 9269228]. doi:10.1016/S0140-6736(97)01164-1

Runge, K., Fiebich, B. L., Kuzior, H., Saliba, S. W., Yousif, N. M., Meixensberger, S., et al. (2021). An observational study investigating cytokine levels in the cerebrospinal fluid of patients with schizophrenia spectrum disorders. *Schizophrenia Res.* 231 (13), 205–213. doi:10.1016/j.schres.2021.03.022

Terryberry, J., Sutjita, M., Shoenfeld, Y., Gilburd, B., Tanne, D., Lorber, M., et al. (1995). Myelin- and microbe-specific antibodies in Guillain-Barre syndrome. *J. Clin. Lab. Anal.* 9 (5), 308–319. [PubMed: 8531012]. doi:10.1002/jcla.1860090506

Uhr, M. (2022). *Bericht zum Ringversuch Gruppe 460, Proteinanalytik*. Duesseldorf, Germany: EQAS Provider INSTAND. Available at: www.instand-ev.de.

Uhr, M. (2024). CSF surveys of INSTAND 2014-2024. Available at: www.instand-ev.de.

Varela, F. J., and Coutinho, A. (1989)). Immune networks: getting on to the real thing. *Res. Immunol.* 140 (9), 837–845. doi:10.1016/0923-2494(89)90043-6

Wahlster, W. (2021). Zehn Jahre Industrie 4.0. Interview FAZ. Available at: www.faz.net/pro/d-economy/digitalisierung-der-produktion-zehn-jahre-industrie-4-0-17267696.html.

Wildemann, B., Oschmann, P., and Reiber, H. (2010). *Laboratory diagnosis in neurology*. Stuttgart: Thieme Verlag.

Wormek, A. (2024). Desktop application as knowledge-based system for cerebrospinal fluid (CSF) analysis. Available at: http://www.wormek.org.

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 11 July 2024
DOI 10.3389/fmolb.2024.1335174

Check for updates

OPEN ACCESS

EDITED BY
Klaus-Peter Hunfeld,
Goethe University, Germany

REVIEWED BY
Jeffrey Budd,
Retired, Saint Paul, MN, United States
Catarina Gregório Martins,
New University of Lisbon, Portugal

*CORRESPONDENCE
T. Keller,
✉ t.keller@acomed-statistik.de

RECEIVED 08 November 2023
ACCEPTED 04 June 2024
PUBLISHED 11 July 2024

CITATION
Seiffert S, Weber S, Sack U and Keller T (2024), Use of logit transformation within statistical analyses of experimental results obtained as proportions: example of method validation experiments and EQA in flow cytometry.
Front. Mol. Biosci. 11:1335174.
doi: 10.3389/fmolb.2024.1335174

Use of logit transformation within statistical analyses of experimental results obtained as proportions: example of method validation experiments and EQA in flow cytometry

S. Seiffert[1], S. Weber[2], U. Sack[1] and T. Keller[2]*

[1]Medical Faculty, Institute of Clinical Immunology, University of Leipzig, Leipzig, Germany, [2]ACOMED statistik, Leipzig, Germany

In laboratory medicine, measurement results are often expressed as proportions of concentrations or counts. These proportions have distinct mathematical properties that can lead to unexpected results when conventional parametric statistical methods are naively applied without due consideration in the analysis of method validation experiments, quality assessments, or clinical studies. In particular, data points near 0% or 100% can lead to misleading analytical conclusions. To avoid these problems, the logit transformation—defined as the natural logarithm of the proportion/(1-proportion)—is used. This transformation produces symmetric distributions centered at zero that extend infinitely in both directions without upper or lower bounds. As a result, parametric statistical methods can be used without introducing bias. Furthermore, homogeneity of variances (HoV) is given. The benefits of this technique are illustrated by two applications: (i) flow cytometry measurement results expressed as proportions and (ii) probabilities derived from multivariable models. In the first case, naive analyses within external quality assessment (EQA) evaluations that lead to inconsistent results are effectively corrected. Second, the transformation eliminates bias and variance heterogeneity, allowing for more effective precision estimation. In summary, the logit transformation ensures unbiased results in statistical analyses. Given the resulting homogeneity of variances, common parametric statistical methods can be implemented, potentially increasing the efficiency of the analysis.

KEYWORDS

external quality assessments (EQA), logit transformation, flow cytometry, method validation, proportions

Introduction

Achieving harmonization and improving the quality of measurements are central goals within the laboratory medicine community. Often, there are a number of measurement methods available for a given measurand from different manufacturers and laboratories. In addition, for methods such as flow cytometry, different experimental settings are used to assess the same measurand. This includes the use of different antibodies and gating strategies that vary from laboratory to laboratory.

In view of this situation, method developers and clinical laboratories are highly motivated to evaluate their measurement methods in internal method validation to gain knowledge about systematic and random errors [Lambert et al., 2020, CLSI guidelines, e.g., EP05]. In addition, external quality assessments (EQAs) are routinely performed.

Method validation typically includes method comparison and precision assessment. It is also important to demonstrate detection capability, robustness to interferences, etc. These experiments are typically analyzed using parametric statistical methods that rely on estimates of the mean and standard deviation (SD) for calculations. However, these methods assume that the data distribution should not deviate significantly from the normal distribution (ND) and that homogeneity of variances (HoV) is maintained over the measurement range, implying that precision is not concentration-dependent. In terms of ND, a visual inspection would be expected to show a symmetrical, bell-shaped distribution with no outliers. In this context, ISO 13528, which guides the design and analysis of EQA, specifies the need for symmetrical distributions.

In the context of EQA, a distinction is made between methods that use reference material and those that do not. If no reference material is available, samples supplied by a service provider are measured by all participating laboratories. The distribution of measurement results is considered, and an *assigned value*—calculated as a robust mean (ISO, 2022; Appendix C, Algorithm A)—is derived directly from the measurements of the participating laboratories. Such assigned values are then used as the basis for individual pass/fail assessments. The establishment of specific pass/fail criteria often involves a balance between observed measurement variability, quality requirements, and the clinical relevance of potential differences. We focus on the common scenario where acceptance criteria are defined by a relative or percentage difference around the *assigned value*.

For example, a typical EQA scenario might assess whether individual laboratories pass if their measurement values for a particular sample are within ± 30 percent of the robust mean (ISO, 2022; Section. 9.3).

The assumptions regarding ND (symmetry) and HoV should be met in such assessments. Although typical laboratory measurands measured in concentrations usually satisfy these assumptions, measurands related to inflammation or tumor incidence are often skewed to the right. In such cases, logarithmic transformation (or, more generally, Box–Cox transformation of the measured values) can help achieve a distribution that does not deviate significantly from the normal distribution and maintains homogeneity of variances. For count data, such as cell counts, square root transformation is often beneficial.

In this context, we consider measurands measured as proportions (0%–100%) or probabilities (0–1) that represent relative measures of specific subgroups within an entity. In flow cytometry, such measurands are exemplified by the evaluation of $CD3^+$ cell subsets, specifically alpha/beta T cells and gamma/delta T cells, both quantified as a percentage of $CD3^+$ cells. Since their combined values add up to 100%, the measured values are inherently correlated.

Another example is the use of probabilities as metrics, such as calculated measurands resulting from multivariable analysis of measured values from a variety of measurands. These probabilities may be generated by logistic regression or other classification methods, such as machine learning or artificial intelligence.

In both scenarios, EQA and combined markers, proportions, and probabilities are bounded between 0% and 100%. Note that in this article, we also use 0 and 1 (representing percentages divided by 100), depending on the context. Because the data are constrained by 0% and 100%, the assumptions regarding ND and HoV are no longer valid, especially for values approaching the limits (<20% and >80%). Consequently, describing, analyzing, or statistically testing experimental data using parametric methods—such as calculating mean and standard deviation, performing *t*-tests and ANOVA for group differences, and estimating variance components (precision) or using ordinary regression techniques—will yield invalid results. This is because these methods assume ND and HoV. In addition, the use of symmetric power limits becomes untenable.

Nevertheless, the naive application of parametric methods to proportions or probabilities is often observed, often due to the convenience of ready-to-use software packages in daily routine and the seemingly straightforward interpretation of results.

Although nonparametric methods could be an alternative, they are less powerful, often require larger sample sizes, and may not provide well-known estimates of bias and precision.

Therefore, we propose to logit transform measured values prior to statistical analysis with parametric methods when measurands are measured as proportions (or probabilities). This article highlights the differences between naive analysis and analysis using logits and provides guidance for interpreting the results.

Materials and methods

EQA data

Simulated data (26 laboratories) from an EQA for alpha/beta T cells (% of $CD3^+$) with an assumed true measured value of 95% and gamma/delta T cells (% of $CD3^+$) with an assumed true measured value of 5% are used. The simulation generates a random number for each laboratory by adding normally distributed noise with an SD of 0.2944, distributed around the assumed value on the logit scale (95% → 2.944). The simulated distribution reflects typical scenarios encountered in the INSTAND program for flow cytometry. In our considerations, we assume that both measurands are highly correlated. For the sake of simplicity, we have assumed a direct relationship between the variables, ensuring that the percentages add up to 100%. Thus, the negative values of measurand A are used as values for measurand B. We have assumed the absence of outliers in this simulation, as outlier detection is beyond the scope of this report.

Precision data

In this simulation study, we generated data representing the results obtained as probabilities from an automated biomarker measurement procedure. Specifically, we measured these probabilities on five different days, with triplicate measurements performed on six different samples. The aim of the experiment is to determine repeatability and overall precision, following the methodology outlined in CLSI (2014).

However, due to the small sample size (15 measurements instead of the recommended 80 measurements for one sample), we chose to estimate repeatability and overall precision pooled over the samples, as suggested by Lambert et al. (2022). This pooling approach assumes the homogeneity of variances.

Within the simulation, the samples were assigned the following values: 4.7%, 14.2%, 37.4%, 64.6%, 85.8%, and 95.3%. The simulation was designed to start with homogeneous variances on a logit scale and then illustrate how the data would be represented on the original probability scale. Within the simulation, the components of imprecision are addressed by adding noise using normal distributed random numbers with the following standard deviations. Within the logit scale, the repeatability (the variability within 1 day) was set at 0.20 for all samples, expressed as the standard deviation. At the same time, the between-day variability was set at 0.12 for all samples. The resulting reproducibility is 0.233 (the square root of the sum of the squared SDs).

To illustrate the impact of using probabilities versus logits, we present variability plots and analyze the precision experiment both naively (per sample) and using logit-transformed values pooled across samples.

The precision components were estimated using random effects ANOVA, as described in CLSI (2014).

Results

Logit transformation

In the logit transformation of a proportion (or probability) p, the following formula is used:

$$\mathrm{logit}(\mathrm{p}) = \ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right), \tag{1}$$

where ln is the natural logarithm.

Table 1 contains selected probabilities and their corresponding logit values that are often encountered in daily work.

TABLE 1 Examples for logit transformation.

p	Percentage (%)	Logit(p)
0.01	1	−4.60
0.05	5	−2.94
0.1	10	−2.20
0.5	50	0
0.9	90	2.20
0.95	95	2.94
0.99	99	4.6

For the back transformation, the following formula is used:

$$\mathrm{p} = \mathrm{antilogit}(\mathrm{x}) = \frac{\exp(\mathrm{x})}{1+\exp(\mathrm{x})}, \tag{2}$$

where x is a value on the logit scale.

Supplementary Material includes an Excel tool for these calculations (Supplementary Material S2).

Application in the context of EQA

In the EQA experiments described above, naive analysis typically involves (i) the calculation of the mean and (ii) the symmetric setting of limits, which is flawed when the underlying distribution is asymmetric. The bias is evident in the calculation of limits, such as the mean ± 30% of the mean, for related percentages, such as 5% and 95%. In a naive analysis, the bounds would be 3.5% . . . 6.5% for the 5% mean and 66.5% . . . 100% for the 95% mean.

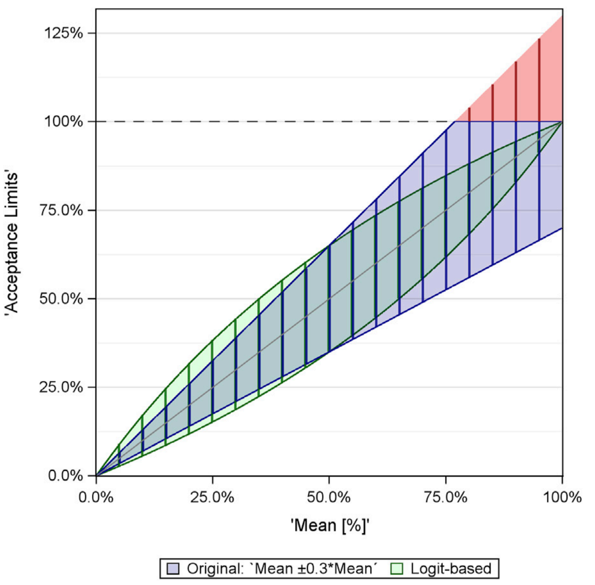


FIGURE 1
Comparison of limits derived within naive EQA analysis (blue/red) and analysis via logit-transformed values (green). The red color refers to the limits >100%, which are not meaningful and are set to 100%.

To calculate accurate limits using the logit scale, the following formula (in conjunction with Eqs 1, 2) is used:

$$\pm Limit_{Logit-based} = antilogit\left(logit(p) \pm logit\left(Limit_{Orig}\right)\right). \tag{3}$$

Figure 1 illustrates the difference between the limits derived from the naive analysis and those obtained using the logit transformation, using the example mean ± 30% x mean. For the 5% and 95% percentages, the ranges between the limits become identical: 3.5% . . . 6.5% for 5% and 93.5% . . . 96.5% for 95%.

Supplementary Material S3 contains an Excel tool for deriving limits on the logit scale.

The following example illustrates the consequences of EQAs. Figure 2 shows simulated data against limits derived naively (left) and after logit transformation of the data and back transformation of the means and limits (right). Supplementary Table S1 lists all values and the EQA results.

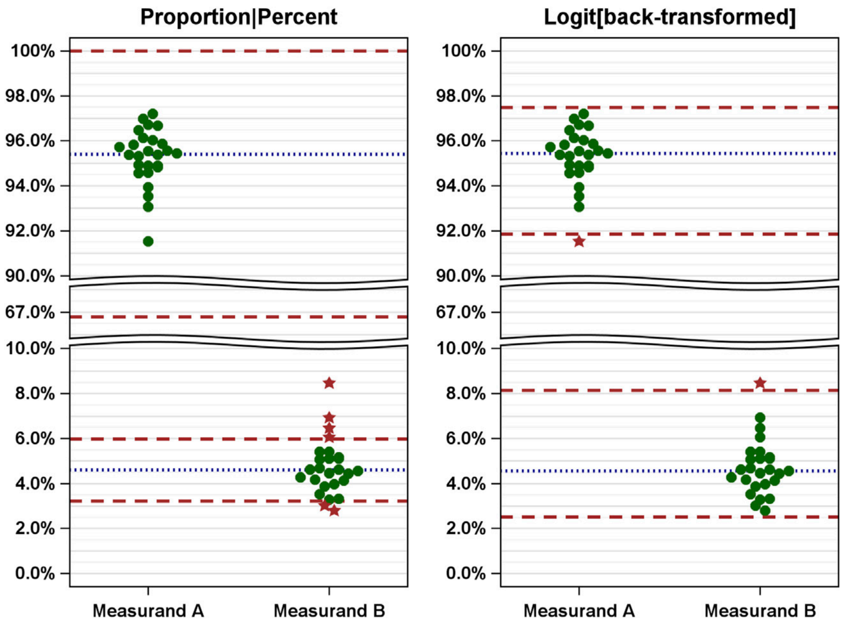


FIGURE 2
Comparison of EQAs (measured values and limits) based on naive analysis (left) and improved analysis (right) using simulated measured results of two measurands [A, alpha/beta T cells (% of CD3⁺); B, gamma/delta T cells (% of CD3⁺)]. The measured values are shown as dots and related to the naively (left: proportions) and correctly (right: logit-scaled values) calculated limits. Green dots, passed; red stars, failed.

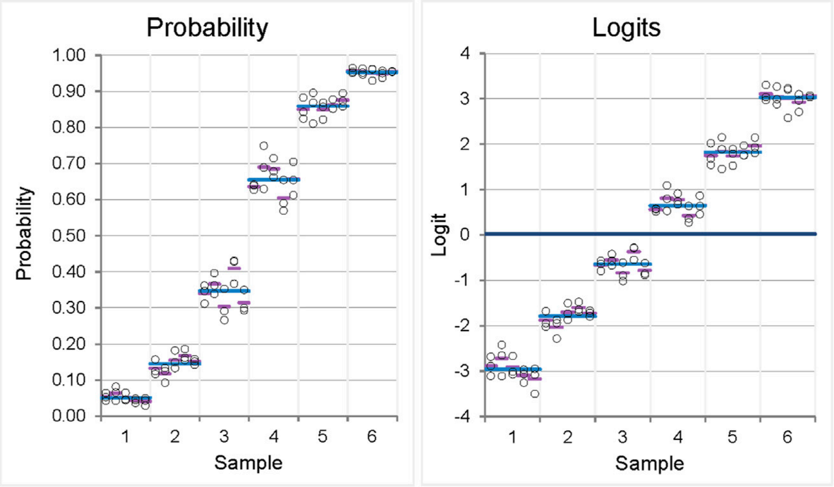


FIGURE 3
Variability plots for originally scaled (left) and logit-scaled (right) measured values of a precision experiment showing homogeneity of variances (homoscedasticity) when logit-transformed values are used for analysis, whereas inhomogeneity of variances is observed when originally scaled proportions are used. The measured values result from a simulation of a 6-sample, ×5-day, ×3-replicates' precision experiment with given values for repeatability and reproducibility on the logit scale (0.20, 0.12). Each circle represents one measurement; pink, mean of the replicates within 1 day; blue, mean of the sample.

The simulation yields skewed distributions with mean values of 95.4% for measurand A and 4.6% for measurand B. In a naive analysis, the limits are between 66.8% and 100% for measurand A and 3.2% and 6.0% for measurand B (Figure 2, left). The span of the ranges defined by these limits varies considerably. Using these limits, all results for measurand A were considered valid, while six results for measurand B were considered invalid.

Conversely, when logit-transformed values are used, the means are nearly identical (see Supplementary Table S1 for details). The limits (calculated according to Eq. 3) are 91.8%–97.5% for measurand A and 2.5%–8.2% for measurand B. Notably, these limits are not symmetric with respect to the mean, but the width of the range defined by the limits is the same (5.7%), as shown in Figure 2 (right). Using these limits, one value for each measurand was identified as invalid.

Application for precision evaluation

In a simulated precision experiment conducted over 5 days, 6 samples were measured in triplicate.

As mentioned earlier, the simulation starts with a given level of repeatability and between-day precision on the logit scale. The associated standard deviations (SDs) in the simulation results show little variation due to random error, as shown in Figure 3 (left). The standard deviations derived from the naive analysis on the probability scale, which represents the scale on which the values would be measured, vary (Figure 3, right). Heterogeneity of the standard deviations is observed, ranging from 0.010 to 0.041 (Table 2, left column; see below for further explanation).

It is important to note that the pooled standard deviations SD_{pooled} derived from logit-transformed values require back transformation. The back-transformed standard deviation based on the logit transformation is now expressed as the lower and upper bounds of the standard deviation. The calculation is performed using the following formula:

$$p \pm SD_{logit-based} = antilogit\big(logit\,(p) \pm SD_{pooled}\big). \tag{4}$$

From Eq. 4, it follows that there are asymmetric "standard uncertainty ranges" [of 1 x SD width] obtained for "above" (see Eq. 5a) and "below" (see Eq. 5b) nominal p ranges.

$$SD_{Logit-based} = antilogit\big(logit\,(p) + SD_{pooled}\big) - p, \tag{5a}$$

$$SD_{Logit-based} = p - antilogit\big(logit\,(p) - SD_{pooled}\big). \tag{5b}$$

Appendix 4 provides an Excel spreadsheet that can be used for these calculations.

Table 2 shows these details of the precision analysis, comparing the naive analysis based on probabilities (left) with the analysis based on logit-scaled scores. In addition, the table shows the back-transformed results. Comparing the SDs from the naive and logit-based analyses at the probability level, it can be seen that the naive analysis leads to biased results.

Most important is the pooled precision across all samples, which is highlighted in the gray cells. This pooled analysis across samples is possible due to the homogeneity of the SDs on the logit scale (as shown in Figure 3, left).

The results of the pooled analysis can be presented in three ways: first, the pooled repeatability and reproducibility are expressed as standard deviations on a logit scale. Second, the back-transformed mean−SD and mean + SD at 50% probability are used. Since the mean−SD and mean + SD are equidistant from 50%, the standard deviation can be presented as a single value. This result can be used as an overall measure of the precision of the measurement procedure. Third, the back transformation of the mean−SD and mean + SD is performed at each sample level; here, the lower and upper SD are presented since the distribution of the measured values is not symmetric.

The results include 95% confidence intervals (CIs), which indicate the uncertainty of the repeatability and reproducibility estimates. The CIs for the pooled analysis are significantly narrower.

The back transformation of the pooled values of repeatability and reproducibility to the logit scale, as performed above for the 50% probability, is possible for any probability. Figure 4 illustrates this calculation. The curves allow the precision for each probability to be determined based on the pooled precision. In Figure 4, the results are compared with the results of the naive analysis, highlighting the discrepancies in the SDs and the wider CIs of the naive analysis.

Discussion

We report on the use of the logit transformation to handle measured values given as proportions or probabilities before applying statistical analyses.

Our results indicate that naive analyses can lead to biased results, especially when the probabilities are close to the limits of 0 (0%) or 1 (100%). In these sub-ranges, the distributions of the values exhibit skewness, as shown in Figure 2 (left). To address these limitations, we suggest using the logit transformation. Analyses are then performed on the logit-transformed probabilities or proportions.

It is important to note that logit transformation is not a novel approach. The literature on predictive models often recommends the use of logit values instead of probabilities (Steyerberg, 2019), such as for model calibration (Ojeda et al., 2023). Furthermore, logit transformation is a special case for transforming data in beta regression, a suggested method for analyzing data observed in (0, 1) intervals (Kischinck and McCullough, 2003; Geissinger et al., 2022).

The use of the logit transformations, among other possible transformations (arcsin (square root), probit, and Fisher's transformation), is also justified by its application in the statistical modeling of binary test results (numbers 0 and 1). One of the most effective methods for this purpose is logistic regression. As internal continuous outcomes, one obtains values on the logit scale, which are then transformed into probabilities to make the results more interpretable. In this context, the application of the logit transformation is straightforward; it is simply the inverse of the transformation. Calculated parameters expressed as probabilities could just result from this or similar classification procedures, so their relationship to logit values, the values calculated in the logistic regression model function, is quite obvious.

It should be noted that differences based on logits can be interpreted as odds ratios, but this is not of interest in the context of this paper, which focuses on the internal and external validation of a measurement method. However, when used in clinical research (e.g., clinical outcome studies), logit transformation provides additional and familiar ways of reporting results.

The use of logit-transformed probabilities is even mentioned in statistical textbooks for medical research (Bland, 2015; Armitage et al., 2015). However, we found no examples of its use in the laboratory medicine literature. This may be because the benefits of the logit transformation are fully realized when the entire range of probabilities, both below and above 50%, as well as near the 0% and 100% limits, is utilized by the measurand. Single measurands typically do not cover this range. However, the two examples in this publication illustrate this scenario. In the case of EQA, the results of the analysis of two corresponding measurands with readings close to 0% for the first and close to 100% for the second become consistent when the simple analysis is replaced by an analysis using the logit-transformed measured values. Another example of measurands that cover the entire measurement range is calculated parameters that reflect classification results based on multiple measurands (Keller et al., 1998; Klocker et al., 2020). Again, the advantage of the logit transformation is highlighted. In addition to providing unbiased estimates of precision, this scale ensures

TABLE 2 Results of precision analysis [repeatability and reproducibility expressed as the standard deviation (SD)/95% confidence intervals (CIs)] calculated by random effects ANOVA as described in CLSI EP05A3. Left: naive analysis per sample on the probability scale; right: analysis pooled across samples on the logit scale, which is possible if homogeneity of variances across samples is given. The results of the pooled analysis are presented on the logit scale and back-transformed to the probability scale for each sample. Since the distribution of the measured values is not symmetric, except for a probability of 0.5, the lower and upper SDs are presented. For a probability of 0.5, only one SD is reported.

Sample	Analysis based on								
	Probability (naïve analysis) analysis per sample			Logit values, pooled analysis					
				Pooled (on the logit scale)		Pooled analysis back-transformed on each sample level			
	Mean	Repeatability 95% CI	Reproducibility 95% CI	Repeatability 95% CI	Reproducibility 95% CI	Repeatability		Reproducibility	
						Lower SD 95% CI	Upper SD 95% CI	Lower SD 95% CI	Upper SD 95% CI
n.a				0.2042 0.1733–0.2486	0.2232 0.1956–0.2728				
1	0.049	0.012 0.009–0.022	0.013 0.010–0.028			0.009 0.007–0.010	0.010 0.009–0.013	0.009 0.008–0.011	0.012 0.010–0.014
2	0.143	0.019 0.013–0.034	0.025 0.019–0.060			0.023 0.020–0.028	0.027 0.023–0.033	0.025 0.022–0.030	0.030 0.026–0.037
3	0.345	0.034 0.024–0.060	0.051 0.037–0.125			0.045 0.038–0.054	0.047 0.040–0.058	0.049 0.043–0.059	0.052 0.045–0.064
n.a	*0.5*			*The pooled results can be back-transformed and reported as the mean ± SD for a probability of 0.5*		0.051 0.043–0.062		0.056 0.049–0.068	
4	0.656	0.041 0.029–0.072	0.049 0.037–0.108			0.047 0.040–0.058	0.045 0.038–0.054	0.052 0.045–0.064	0.048 0.043–0.059
5	0.861	0.028 0.020–0.049	0.028 0.023–0.046			0.026 0.022–0.033	0.023 0.019–0.027	0.029 0.025–0.036	0.025 0.022–0.030
6	0.954	0.010 0.007–0.018	0.010 0.008–0.017			0.010 0.008–0.012	0.008 0.007–0.010	0.011 0.009–0.014	0.009 0.008–0.011

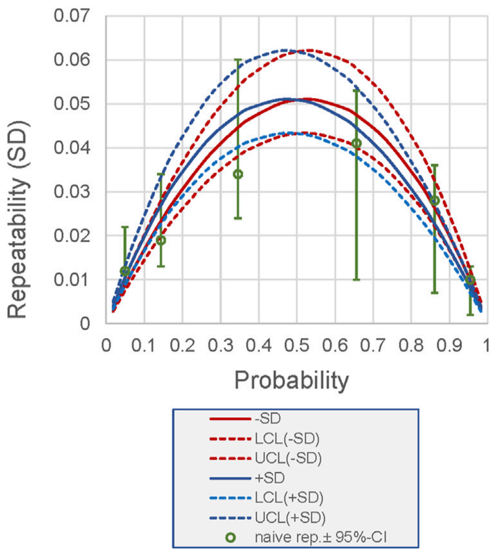


FIGURE 4
Relationship between probability (as a measurement result of a measurand measured as a proportion or probability) and its repeatability for the example shown in Figure 3. The repeatability is derived from the pooled analysis using logit-scaled values. Back-transformed values are presented. Lines, back-transformed SDs (+, −) derived from pooled repeatability; dashed lines, 95% CI. Dots, sample-specific repeatability derived from naive analysis (which does not allow pooled estimation), including 95% CI. LCL, lower confidence limit; UCL, upper confidence limit.

the homogeneity of variances, i.e., the random error does not systematically depend on the measured value. This allows the use of parametric statistical methods such as regression, ANOVA, or even pooling of precision across samples. The latter leads to a significant increase in statistical power or a reduction in the sample size required for the precision experiment.

Finally, we demonstrate the methodology on two examples from these areas: EQA of flow cytometry data measured as proportions and a method validation experiment (precision) on values expressed as probabilities. Figure 4 summarizes the advantages of logit-based precision analysis: it provides unbiased and more precise estimates of the precision components.

It is important to note that the analysis of other method validation experiments, such as method comparisons, can benefit significantly from our proposed approach.

Instead of using real data, we chose to use simulated data. Although this may be seen as a limitation of our study, it is important to note that these simulations were directly inspired by real examples from our daily work. In addition, we initially encountered naive analyses that had previously been used for statistical evaluations in the EQA program. Based on the considerations reported here, the EQA-related analyses are currently modified. In the case of the precision data of a calculated parameter, the precision is often found to be suboptimal and not meet the acceptance criteria. In that case, the variances of the contributing measurands add up due to error propagation. It is not the purpose of this article to report on these specific problematic data sets.

Moreover, real-world data often present their own challenges. In addition to the effect to be demonstrated, there may be other issues such as outliers, imperfect correlations, or dilution of effects due to simple imprecision. These complications require a discussion of all side effects, which can sometimes overshadow the main effects.

For these reasons, we decided to use simulated data.

When data transformations are applied, a thorough discussion of the back transformation is required. Although an unbiased analysis with the ability to use all the tools of parametric statistics is advantageous, it often comes at the cost of more complex, or at least unusual, handling of the results. For example, when standard deviations of proportions are evaluated, the results are more complex and require additional explanation. One option is to present the results on a logit scale, which may not be practical: this would require the user to be able to conceptualize in logits, which may not be a realistic expectation. However, when different values expressed as proportions are evaluated in parallel, it may be useful to compare the results directly on the logit scale.

Another option is to back-transform the results. It is important to note that only points (e.g., mean +SD and mean − SD) can be back-transformed, not a range such as SD. Due to the skewed distribution of the underlying values (proportions and probabilities), the resulting mean ± SD will also be asymmetric. Reporting the precision for a 50% probability as an abstract but easy-to-read overall measure is then advantageous because it is symmetric at approximately 50%.

Finally, we strongly recommend the use of logit-transformed data in statistical analyses of clinical laboratory and quality control data when the measures are proportions or probabilities. This approach enhances the interpretability and power of the results, thereby facilitating their application in the relevant fields.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

SS: conceptualization, data curation, funding acquisition, investigation, methodology, project administration, supervision, validation, writing–original draft, resources, and writing–review and editing. SW: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, writing–review and editing, and writing–original draft. US: conceptualization, data curation, funding acquisition, investigation, methodology, project administration, supervision, writing–review and editing, and resources. TK: conceptualization, data curation, investigation, methodology, project administration, writing–review and editing, validation, formal analysis, funding acquisition, software, supervision, visualization, and writing–original draft.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article.

The funds required for the publication and parts of the underlying research were provided by the Society for Promotion of Quality Assurance Medical Laboratories (INSTAND e.V.), Düsseldorf.

Conflict of interest

Author TK is the owner of the company ACOMED statistik. Author SW was employed by the company ACOMED statistik.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2024.1335174/full#supplementary-material

References

Armitage, P., Berry, G., and Matthews, J. N. S. (2015). *Statistical methods in medical research.* John Wiley and Sons.

Bland, M. (2015). *An introduction to medical statistics.* 4th ed. Oxford University Press.

CLSI (2014). *Evaluation of precision of quantitative measurement procedures; approved guideline – third edition. CLSI document EP05-A3.* Wayne, PA: Clinical and Laboratory Standards Institute.

Geissinger, E. A., Khoo, C. L. L., Richmond, C. R., Faulkner, S. J. M., and Schneider, D. C. (2022). A case for beta regression in the natural Sciences. *Ecosphere* 13, e3940. doi:10.1002/ecs2.3940

ISO (2022). *Statistical Methods for use in proficiency testing by interlaboratory comparison.* ISO 13528:2015, corrected version 2016-10-15, published 2020-09).

Keller, T., Bitterlich, N., Hilfenhaus, S., Bigl, H., Löser, T., and Leonhardt, P. (1998). Tumour markers in the diagnosis of bronchial carcinoma: new options using fuzzy logic based tumour marker profiles. *J. Cancer Res. Clin. Oncol.* 124, 565–574. doi:10.1007/s004320050216

Kieschnick, R., and McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Stat. Model.* 3, 193–213. doi:10.1191/1471082x03st053oa

Klocker, H., Golding, B., Weber, S., Steiner, E., Tennstedt, P., Keller, T., et al. (2020). Development and validation of a novel multivariate risk score to guide biopsy decision for the diagnosis of clinically significant prostate cancer. *BJUI Compass* 1, 15–20. doi:10.1002/bco2.8

Lambert, C., Demirel, G. Y., Keller, T., Preijers, F., Psarra, K., Schiemann, M., et al. (2020). Flow cytometric analyses of lymphocyte markers in immune oncology: a comprehensive guidance for validation practice according to laws and standards. *Front. Immunol.* 11 (2169), 1–20. doi:10.3389/fimmu.2020.02169

Ojeda, F. M., Jansen, M. L., Thiéry, A., Blankenberg, S., Weimar, C., Schmid, M., et al. (2023). Calibrating machine learning approaches for probability estimation: a comprehensive comparison. *Stat. Med.* 42, 5451–5478. doi:10.1002/sim.9921

Steyerberg, E. W. (2019). *Clinical prediction models: a practical approach to development, validation, and updating.* 2nd ed. Cham: Springer.

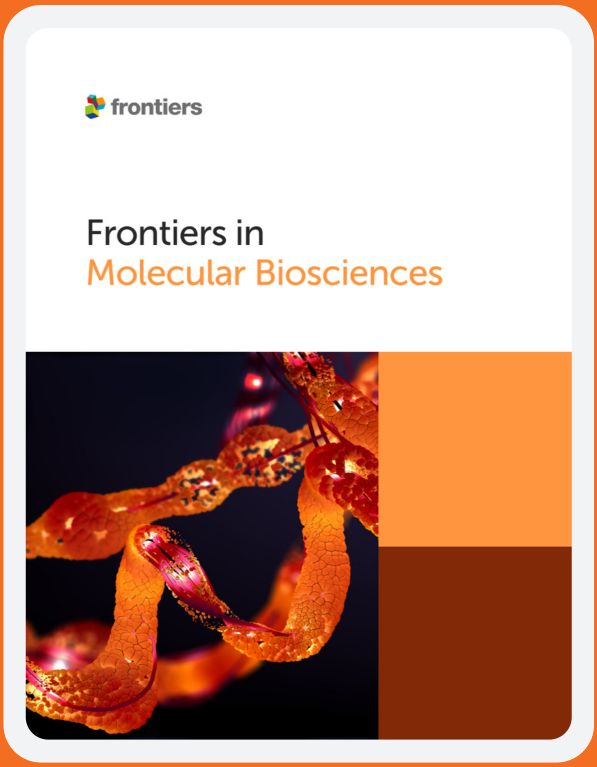
Frontiers in
Molecular Biosciences

Explores biological processes in living organisms on a molecular scale

Focuses on the molecular mechanisms underpinning and regulating biological processes in organisms across all branches of life.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Molecular Biosciences

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact