Standardization of protein biomarker measurements: Is it feasible?

HEINZ SCHIMMEL, INGRID ZEGERS & HENDRIK EMONS

European Commission, Joint Research Centre, Institute for Reference Materials and Measurements, Geel, Belgium

Abstract

The standardisation of measurements of protein biomarkers, which are potentially heterogeneous in terms of fragmentation, modification, substitution, primary, secondary, tertiary and quaternary structure, is a demanding task. However, they are a prime target for standardisation efforts due to the importance of protein biomarkers in diagnostics and health care and the typically observed significant discrepancies in measurement results obtained with non-standardized platforms. Based on the experience gathered during successfully completed projects for the production of reference materials, pragmatic approaches are described how standardisation could become feasible despite the fuzziness of the target analytes.

Key Words: Certified Reference Materials, calibration, metrological traceability, commutability, quantitative profiling, measurement harmonisation

Introduction

The scientific progress in clinical chemistry and advancements of the analytical capabilities to measure an increasing number of parameters for human diagnostics have created an enormous amount of data. Their reliability is crucial for health-related decisions and affects finally the whole society. In addition, the worldwide mobility of people, demanding the equivalence of diagnostic findings for an individual at the global scale, and the globalisation of the production and distribution of diagnostic instrumentation, test kits and reagents require at least comparable analytical results in space and time.

The issue of standardisation for measurements in clinical chemistry has been taken up in legislation and in international standards. The metrological principles have been described, e.g., in the standards issued by the International Organization for Standardization (ISO), namely ISO/CEN 17511 and ISO 18153. Since 1998 the EU Directive on In Vitro Diagnostic Medical Devices (IVD-MD) (Directive 98/79/EC) requires traceability of calibrators and control materials to reference measurement procedures and/or reference materials of higher order. This legislation has a worldwide impact, as IVD manufacturers must ensure that the systems they market have been properly calibrated against certified reference materials and reference measurement procedures.

The underlying concept is based on the principle that metrological traceability to a stable reference, i.e. the SI or a stable materialized standard carrying the measured quantity, via an unbroken chain of unbiased comparisons with known uncertainty will guarantee comparability of measurement results. Consequently metrological traceability has to be considered as a means to achieve comparability and not as a formal purpose on its own. For macromolecules used as clinical markers such as proteins, the establishment of a meaningful and efficient traceability chain is particularly demanding. A “classical” monoparametric approach from analytical chemistry, directed only to the identification and quantification of a single type of molecules or parts thereof, does not provide all relevant information required for the standardisation of the methods and procedures (platforms) used in a routine laboratory setting. It is rather the measurement of a specified functionality of the target molecule which matters. The characterisation of the amount of substance (and even more of the mass of the protein, also when only present in a single form) alone would usually not sufficiently describe the properties of the macromolecule and the higher order standard well enough. Therefore all relevant influence parameters and quantities need to be controlled and characterised to remove the arbitrary character of the calibration, i.e. to avoid...
uncontrolled calibration bias, and consequently to increase the reproducibility of the calibration process. Knowledge of the relevant influence parameters and quantities will also allow the reproducible production of reference materials used for calibration and having only a limited lifetime. Only under such conditions the comparability of measurement results traceable to the SI or other stated references can be achieved over space and long time periods. The complex, macromolecular nature of the targets creates considerable metrological challenges and the traceability concept is unlikely to be monoparametric. Only a minor fraction of the several hundred clinically relevant analytes (~15%) are well defined molecules or could be standardised with internationally accepted reference methods which are defining the measurand. The vast majority of the clinically relevant parameters belong to the group of complex macromolecules and most of them are proteins.

The laboratory medicine community is facing quite some difficulties with lacking comparability of measurement results obtained with the same platform over time or between platforms. In external quality assurance schemes like the ones organised by UK-NEQAS or by the Deutsche Vereinte Gesellschaft für Klinische Chemie und Laboratoriumsmedizin (DGKL) the coefficients of variation for results on protein biomarker measurands range from 5–6% for very well determined systems like albumin up to 25–50% or even higher for the complex proteins like troponin I or the brain natriuretic peptide BNP.

Improvements could be achieved, for instance, through IFCC by advocating for highly precise reference methods which are based on thorough, state of the art biochemical investigations, and which define de facto the measurands. Reference systems could be established by combining those reference methods with networks of experienced laboratories and reference materials, whose production is also supported by IFCC. They have significantly improved the comparability of laboratory medicine data (e.g. for enzymatic activity, serum proteins, HbA1c, etc.).

The high variations found for a large number of analytes is problematic because results from different laboratories and methods cannot be compared directly, and it is in many cases not possible to define common reference intervals or cut-off points for diagnostic decisions. This could also mean that the diagnostic power of some biomarkers cannot be exploited to the full extent.

Based on experience gathered during protein measurement standardisation projects this paper aims to highlight current metrological concepts for achieving international equivalence of measurement results on macromolecular biomarkers of clinical relevance and point to some critical aspects in the standardisation process.

General considerations for a standardisation approach

The comparability of analytical data is based on three basic concepts of metrology: definition of the measurand of interest, adequate estimation of the measurement uncertainty as part of the measurement result and the establishment of the metrological traceability for the results.

Standardisation of proteins and the preparation of sufficiently characterised certified reference materials, suitable for the calibration of routine methods and to guarantee a stable measurement scale over long time, i.e. also beyond the lifetime of a particular certified reference material, is a multi-step and often iterative process. The impact of different potential influence properties, such as fragmentation, modification (glycosylation, oxidation, deamination etc.), secondary and tertiary structure, multimeric stage, degree of aggregation, and quantities (buffer composition, quantitative isoform profile) would have to be identified and, in case of quantities, to be quantified.

The full characterisation of the relevant properties of a protein, in particular in matrices and when present at low concentrations, cannot be achieved with the currently available technical means. However, experiments and verification steps should be designed in a way that critical parameters can be identified and consequently be kept under control. The certified reference material should have characteristics with respect to the relevant influence properties and quantities that are similar to those of typical routine samples. Otherwise the calibration could be biased due to analytical artefacts. Therefore the assessment of the impact of the raw material selection, material processing steps and methodological changes are keys to be able to come up with a truly standardised and reproducible measurement system.

Macromolecules like proteins, eventually natively represented by a mixture of isoforms, are typically quantified by measuring the amount of substance of a certain part of the sum of isoforms or particular isoforms (epitope for immunoassays). Insisting on a very specific and strictly species-limited definition of the measurand, such as the amount of substance of a certain epitope in certain isoforms having certain structures, would mean that the measurand of routine assays targeting different epitopes would be different and therefore by definition not comparable. This would fix the status quo of non-comparable measurement results, which would not be in the interest of those having to interpret the discrepant measurement results. In many cases different measurands/analytes are in a sufficiently constant relationship in typical routine samples providing a chance for standardisation to an acceptable level of comparability even though the actual measurand might be different.
This principle was applied for the IFCC reference system for glycated haemoglobin as represented by HbA1c, which occurs quite constantly at a level of about 80% of all glycated haemoglobins in human blood [3]. Some harmonisation might be achieved in cases where the different measurable targets are not in a relatively constant relationship in patient samples. However, tolerable levels of discrepancy between measurement results might be smaller due to clinical needs. In such cases the clinically most relevant isoform or analytical target (i.e. epitope or protein region) would have to be determined or a consensus on a common analytical target would have to be reached. This would also imply that some routine methods, which do not yet measure (or measure only to an insufficient extent) the clinically most relevant or consensus target, would have to be modified accordingly. The IFCC working groups and committees dealing with the standardisation of human chorionic gonadotropin (hCG) [4] and troponine I (TnI) [5] have used this approach.

Before engaging into the standardisation of a biomarker it is advisable to assess first the consistency of the entities measured by the different methods to be standardised. Sound decisions on the way forward can only be made on the basis of the outcome of such method comparisons.

Initial method comparison

The results of external quality assurance schemes (EQAS) are a very useful source of information on the comparability of routine methods and may indicate groups of routine methods with similar specificity. Therefore, they can already point to sources of discrepancies between routine methods. However, the EQAS schemes are mostly using pooled samples, therefore differences between individual samples may be averaged out.

Hence, ideally the first step in the standardisation of protein biomarkers would be a kind of commutability study, i.e. using a larger number of fresh, unprocessed samples from healthy individuals and patients in a pair-wise comparison of routine methods [6]. Already at this step some candidate reference materials in different formats could be measured in parallel with the samples from patients and healthy individuals. Clear preference should be given to the use of samples from individuals rather than pooled samples, since pooled samples bear the risk that pronounced differences between individuals (whose samples are finally also measured by the routine methods) may be averaged out, therefore not allowing the proper assessment of the specificity of the routine method.

A lacking correlation between the pair of routine methods would reveal they have different specificities. Using a calibration material whose measurement result would fall onto the interpolated correlation line for both methods in such a situation would mean that only the averages over a given sample population could be harmonised. In many cases this may already have a positive effect on the comparability of results from different assays. However, depending on the magnitude of the data scatter around the correlation line, pronounced differences between results on an individual patient sample may persist. Even after recalibration with the commutable calibrator discrepancies may stay at an undesirable or even worse, clinically non-tolerable level. Since comparability of results on the majority or ideally on all patient samples is the ultimate goal of standardisation, the question would arise in the latter case whether the production of a higher-order reference material would make sense at this stage. The more appropriate approach would be to agree on a common target (measurand) which could be defined as a chemical entity or via a reference measurement procedure (of high metrological quality and robustness), taking information on its clinical utility into consideration. At least one of the discrepant routine methods would then have to be adapted or modified with the aim to produce results for the agreed measurand or which correlate more closely with the reference measurement procedure.

Non-linearity of the correlation line would be a major issue for standardisation and may indicate a change of the protein isoform pattern with the health status of the individuals. It may restrict the standardisation effect to a certain concentration interval for which the correlation between methods is linear. On the other hand such an observation provides valuable information on the role of certain isoforms as a more specific biomarker for the health status.

If the pair-wise compared methods correlate well and the correlation is linear across the relevant concentration interval there is a good chance for successful standardisation.

As mentioned above candidate reference material preparations (pilot batches) could be analysed in parallel to the fresh samples from healthy subjects and patients. If certain formats appear commutable they might be suitable candidates for higher-order reference materials.

Lacking commutability (not following the same correlation between two methods as fresh samples from healthy subjects and patients) for a candidate reference material can have various reasons. For protein measurements the matrix interactions play a major role in their behaviour in analytical systems. Hence non-commutability could be caused by differences in the matrix influence between candidate reference material and routine samples. This problem can be overcome by changing the reference material presentation bringing the matrix properties closer to a routine sample.

The non-commutability of the reference material could also be explained by differences in method specificity. It might be that the methods correlate
very well on routine samples due to a relatively constant relationship of the different targets detected by the different methods. If a calibrator turns out not to be commutable it could be besides matrix effects also due to the occurrence of atypical isoforms in the calibrator which are differently detected by the routine methods and therefore bias at least one of them. Consequently, although correlating well for fresh routine samples, different results would be observed for the non-commutable reference material. This problem could be overcome by changing the format and processing of the calibrator and could be supported by changes in the method to make them more robust towards such isoform changes.

An example for atypical isoforms possibly occurring in candidate reference materials and their negative impact on the commutability of reference materials is ceruloplasmin. It became obvious during the characterisation of ERM-DA470k/IFCC [7], calibrated against ERM-DA470/IFCC [8] that there are two groups of routine methods for which results differ by approximately 10% (Figure 1). A difference of the same magnitude and direction between those two groups of routine methods can also be observed in EQAS schemes (data not shown). However, all routine methods included in the study were calibrated against ERM-DA470/IFCC. Looking closer to the reasons for these discrepancies revealed that an atypical isoform of ceruloplasmin, normally not present in native serum samples, has been detected in ERM-DA470/IFCC (by immunoblotting, data not shown). This atypical isoform is seemingly detected by the two groups of optimised routine methods to a different degree. Within these two groups of methods ERM-DA470/IFCC appears commutable, but not amongst the two groups. The properties of ERM-DA470k/IFCC with regard to ceruloplasmin resemble more a native serum sample. Hence one quantity, namely the fraction of the atypical isoform influencing the measurement results, has not been sufficiently characterised in the older ERM-DA470/IFCC, leading for at least one group of routine methods to an uncontrolled calibration bias. Hence ERM-DA470/IFCC was unsuited as a calibrator for the value assignment to ERM-DA470k/IFCC and the protein could not be certified in ERM-DA470k/IFCC so far. Research is ongoing to resolve this problem. It should be noted at this point that a value assignment on the basis of mass spectrometry only, would not have recognised this problem and would have left a large degree of arbitrariness in the calibration of routine methods and hence would not have standardised the ceruloplasmin measurements. In fact the measurand would have changed through the traceability chain and the routine measurement systems would only be partially calibrated.

In summary trying to advance the understanding on the reasons for discrepancies between routine methods and the non-commutability of certain candidate reference materials is increasing the understanding of the performance of routine methods. It will give valuable information on the influence parameters and the standardisation concept to be applied, will allow to take corrective actions in terms of reference material preparation and routine method design, and will also provide useful information for the implementation of reference systems, including hints on the proper use of the reference materials. Consequently it increases the efficiency of standardisation efforts.

Figure 1. A: Results from value assignment data on ceruloplasmin. The transfer factors shown are the [CER]ERM-DA470k/IFCC/[CER] ERM-DA470. Different symbols represent results obtained with different methods. B: Table of results from commutability studies for ERM-DA470 for the six different methods (numbered 1 to 6) represented in the value assignment data. “1” means that the value for ERM-DA470 is within the 95% prediction interval for patient data for the two methods, a “0” means it that the value for ERM-DA470 is outside the prediction interval.
Reference material preparation
The preparation of a certified reference material is a relatively complex process which typically requires a feasibility study directed to understand the relevant properties of different formats of pilot batches. Typically such pilot batches would be included in the method comparison exercise to evaluate which presentation of a reference material would likely be commutable, which is one condition to serve as an adequate calibrator for routine methods. Commutability of a reference material is an indicator of the potential degree of harmonisation achievable for routine methods through calibration with the reference material. The reference material used for the calibration of routine measurement procedures must behave similar to routine samples, must be stable over its lifetime and has to be homogeneous at the amount of sample used for analysis. All relevant influence parameters need to be controlled and quantified to ensure reproducibility of the production and consistency of the measurement scale over time.

This can only be achieved by various dedicated studies evaluating the impact of the processing steps on the properties of the reference material. It has to be ensured that material alterations do not cause a calibration bias.

The importance of understanding the impact of processing steps on the properties of the reference material is illustrated, for instance, by an observation related to the C-reactive protein (CRP) during the production of ERM-DA470k/IFCC [7]. This is a lyophilised material, which results in certain advantages in terms of storage and distribution logistics as well as potentially on the material stability. After lyophilisation the amount of detectable CRP in the reference material dropped by about 20% (Figure 2). The lyophilisation step did not have a significant impact on the detectable amount of the other 12 proteins which were certified in ERM-DA470k/IFCC. It was therefore decided not to certify the CRP concentration in ERM-DA470k/IFCC and to produce a liquid frozen material (ERM-DA472) for this protein [9].

This finding emphasises the importance of evaluating the impact of the different material processing steps (including the initial sampling of the raw material) on the properties and suitability of a reference material. Moreover, it is a striking example for a reference material which appears to be perfectly commutable, but could create nevertheless a major calibration bias, depending on the actual definition of the measurand and the value assignment process applied. Such observations related to apparently commutable reference materials are becoming more likely for measurements where there is already a high degree of harmonisation between methods as it is the case for CRP. Again, a value assignment based on the mass or amount of substance only would not have characterised the CRP protein in the different reference material presentations comprehensively enough to avoid arbitrariness.

Use of the reference material and implementation of reference systems
According to international standards (ISO Guide 31 [10], ISO Guide 33 [11], ISO Guide 34 [12], ISO 15194 [13]) a reference material producer would also have to provide instructions on the proper use of the calibrator. In fact the reconstitution and conditioning of a reference material before the actual measurement may have a significant impact on the measurement results. This is not only related to the accuracy of performing the reconstitution protocol in terms of analyte concentration, but also to a potential

![Figure 2. Results from value assignment measurements for CRP in ERM-DA470k/IFCC (▲) and for ERM-DA472/IFCC (●). The transfer factors shown are [CRP]ERM-DA470k/IFCC/[CRP]ERM-DA470 and [CRP]ERM-DA472/IFCC/[CRP]ERM-DA470. The materials are lyophilised and liquid frozen, respectively. The plot shows the mean values of individual datasets and the standard deviation.](image-url)
time course related to reaching a stable material state similar to a typical patient sample in terms of structure, aggregation etc.

According to the documentary standards mentioned above, a reference material producer would have to provide also known information on the commutability and possible shortcomings of the reference material if used for the specified purpose.

The effect of this step on the standardisation is rarely systematically investigated but is in the view of the authors an important part of a standardisation concept/reference system and should not be neglected.

It may also be useful to develop a harmonised calibration protocol to further improve inter-method comparability. Although a commonly used higher order reference material may be commutable differences in details of the calibration procedure for routine methods, such as number and level of calibration points or type and properties of diluent used, may still contribute to relevant differences in measurement results. Therefore guidelines on the calibration, if necessary including detailed experimental protocols such as those recently published for serum proteins [14], could improve the situation.

Conclusions

The standardisation of measurement of protein biomarkers is possible provided that all parties involved in the standardisation and calibration process are aware of the relevant scientific and technical issues and the potential pitfalls. Metrological traceability to higher order reference materials and reference methods should not be considered as formalism but rather as an efficient tool to achieve standardisation of biomarker measurements. Establishing an appropriate and efficient traceability chain requires alertness with regard to critical parameters influencing the measurement results. These parameters have to be quantified and controlled to be able to maintain a stable measurement scale over a longer time, i.e. to achieve true standardisation. Otherwise the claimed standardisation may turn out to be at best a temporary harmonisation approach. For reference materials only small changes in the processing may have a pronounced impact on the measurement results and would consequently lead to a biased calibration. Hence a careful evaluation of the impact of different reference material formats and processing steps on the properties of the reference material and its suitability for calibration (such as, but not limited to, commutability) is a crucial requirement.

Standardisation of protein biomarker measurements is certainly a demanding task. However, following a concept as described in this article would in many cases allow the implementation of a standardisation approach with justifiable efforts, in particular for high priority and high volume biomarker measurements.

It is important to understand the reasons for unexpected observations and discrepancies in the measurement process. By that, parameters influencing the measurement results can be identified and the relationship between different analytes targeted by the different routine measurement procedures can be understood.

The presented examples demonstrate that there are hardly any generally applicable concepts for standardisation and reference material production for protein biomarkers. They have frequently their own specific peculiarities due to their complexity and heterogeneity.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Feasibility of protein standardisation

33


