Harmonization of blood-based indicators of iron status: making the hard work matter

Andrew N Hoofnagle

Departments of Laboratory Medicine and Medicine, University of Washington, Seattle, WA

ABSTRACT

Blood-based indicators that are used in the assessment of iron status are assumed to be accurate. In practice, inaccuracies in these measurements exist and stem from bias and variability. For example, the analytic variability of serum ferritin measurements across laboratories is very high (>15%), which increases the rate of misclassification in clinical and epidemiologic studies. The procedures that are used in laboratory medicine to minimize bias and variability could be used effectively in clinical research studies, particularly in the evaluation of iron deficiency and its associated anemia in pregnancy and early childhood and in characterizing states of iron repletion and excess. The harmonization and standardization of traditional and novel bioindicators of iron status will allow results from clinical studies to be more meaningfully translated into clinical practice by providing a firm foundation for clinical laboratories to set appropriate cutoffs. In addition, proficiency testing monitors the performance of the methods over time. It is important that measures of iron status be evaluated, validated, and performed in a manner that is consistent with standard procedures in laboratory medicine. *Am J Clin Nutr* 2017;106(Suppl):1615S–9S.

Keywords: assay, bias, clinical bioindicator, harmonization, imprecision, iron status, reference material, reference method, standardization

INTRODUCTION

The translation of clinical research relies on reproducible, interpretable results (1). When blood-based bioindicators are used to categorize iron status, they are assumed to be accurate. The ability to achieve a desired level of accuracy hinges on the quality of the assay and the approach to calibration of the assay signal into a concentration. Concentrations that are measured during clinical research studies should be meaningful with respect to the concentrations measured after assays are deployed in clinical settings. This method would allow for conclusions from clinical research studies to be directly translated to clinical care with the goal that bioindicator concentrations are correlated with outcomes and treatment decisions in meaningful ways.

This paper was prepared to support discussions that were part of an NIH workshop that was focused on bioindicators of iron status and risks and benefits of the iron supplementation of pregnant women and young children. Although most research has focused on circulating bioindicators of iron status that have traditionally centered on identifying iron deficiency (ID) and iron-deficiency anemia, the workshop included discussions of bioindicators of the iron-replete or iron-adequate state. Furthermore, the workshop addressed issues related to the many different bioindicators of iron status that are available, their strengths and weaknesses, and their propensities to reflect the different stages of iron status. An overarching question, as specified by recent reports from the US Preventive Services Task Force (2, 3), is the extent to which these hematologic measures can be linked to meaningful health outcomes. Other presentations discussed the methods used and the challenges in interpreting blood-based indicators of iron status, particularly in the context of inflammation (4–6). This paper focuses on the procedures that are used in laboratory medicine to minimize bias and variability, namely harmonization, standardization, and proficiency testing. In addition, immediate steps that would help translate research studies of iron status into improvements in population health are summarized.

ACCURACY OF MEASURES

The term accuracy has many connotations. From an analytic chemistry perspective, accuracy can be defined as the closeness in agreement between the test result and the true result (7). For each measurement in each sample, there are the 2 following important contributions to inaccuracy: bias and imprecision.

Analytic bias

Analytic bias is the systematic error of the measurement. There are the following 2 sources of bias in any one measurement: *1*) the bias from the calibration system and *2*) the bias from the individual sample itself. A calibration system is used to transform the assay signal into a concentration. Although there are

Presented at the workshop "Iron Screening and Supplementation in Iron-Replete Pregnant Women and Young Children" held by the NIH Office of Dietary Supplements, Bethesda, MD, 28–29 September 2016.

Supported in part by the University of Washington Nutrition and Obesity Research Center (grant P30 DK035816; to ANH).

Address correspondence to ANH (e-mail: ahoof@u.washington.edu).

Abbreviations used: CAP, College of American Pathologists; ID, iron deficiency; ISO, International Organization for Standardization; SF, serum ferritin; sTfR, soluble transferrin receptor.

First published online October 25, 2017; doi: https://doi.org/10.3945/ajcn.117.155895.

many ways to do this, the most robust way may be the use of calibrators that are made from the actual sample type that will be analyzed in the assay (this is called matrix matching). An example relevant to nutrition is the use of 25-hydroxyvitamin D calibrators that are made in human serum, in which human serum is stripped of 25-hydroxyvitamin D with the use of a charcoal filter, and 25-hydroxyvitamin D is added back into the stripped serum at defined concentrations. Note that serum and plasma are different sample types, and not all assays will perform the same in serum and in plasma (8). Other types of calibrants include purified analyte or purified analyte spiked into an artificial mixture of proteins (e.g., bovine serum albumin and bovine immunoglobulins in buffered saline). The latter is the most common form of calibrant used in automated clinical chemistry and clinical immunochemistry platforms. Regardless of the calibration system used, the assay signal from the calibrators must be similar to the signal from the human samples that will be analyzed by the assay that have a similar concentration of analyte.

The other source of analytic bias (i.e., from the individual sample itself) is specific to each sample. These sample-specific biases, also known as matrix effects, can be very difficult to predict a priori. Matrix effects change the relation between the signal that is generated in an assay and the amount of analyte in the sample. In other words, the slope of a calibration curve that is made in certain human samples may be very different from the slope of the calibration curve that is used in the assay. An example of a matrix effect is the error observed in an assay that measures the amount of light passing through a sample (e.g., a turbidimetric assay) because of chylomicronemia, which causes light to scatter as it passes through the sample. Unfortunately, it is only truly possible to identify sample-specific matrix effects via spike-recovery approaches. Although examples of these analyses exist (9, 10), they are infrequently performed in clinical and clinical research laboratories because of the complexity of the analytic processes.

Sample-specific biases are specific to the platform being used to make the measurements (e.g., Beckman DxI, Roche Cobas, and Abbott Architect). This issue is particularly true for protein immunoassays that use different reagent antibodies to recognize different epitopes from the same protein. Intellectual property and marketing concerns typically drive the development of divergent reagents in the in vitro diagnostic industry. The use of different reagent antibodies leads to different matrix-specific interferences because different molecules will cross-react with the different reagent antibodies, thus leading to variable assay signal on each analytic platform from the same amount of analyte in a specific human sample. For nonprotein analytes such as iron, sample-specific matrix effects are less common but possible (e.g., chelators).

Analytic imprecision

The second component of inaccuracy is the analytic imprecision or variability of the assay. This variability includes within-day variability and between-day variability, which can be approximated straightforwardly with the use of repeated measures (i.e., the assay is run many times on the same day, or the assay is run once on many different days, respectively). In practice, clinical laboratories include quality-control samples in each batch or throughout the day to assess the performance of the assay. Unexpected variability results in troubleshooting and corrective action (e.g., instrument maintenance and recalibration). There are many other sources of analytic variability that can affect the performance of bioindicators in the clinical care of patients. For example, between-instrument variability and between-laboratory variability as well as between-operator variability or between-reagent lot variability can be relevant. These issues are more difficult to assess, but proficiency testing can help provide some data to estimate these effects as discussed below (see Proficiency Testing).

In clinical care and clinical research, a single measurement on a single sample that is drawn at a single point in time and measured in a single laboratory on a single platform is often used to establish iron status. As a result, the intraindividual variability of the bioindicator concentration is also compounded into the accuracy of the clinical measurement. Note that this relation affects the clinical accuracy of the measurement rather than the analytic accuracy. Intraindividual variability can occur because of actual variations of bioindicator concentration during the day, week, or month, but also variations in the blood draw technique and sample processing in the clinical laboratory (e.g., centrifuge speed, centrifuge temperature, time between sample collection, and separation from cells) (11). In some instances, it is appropriate to repeat measures in clinical care to avoid making a diagnostic error (e.g., in the diagnosis of diabetes or in screening for prostate cancer).

HARMONIZATION OF MEASURES: REMOVING ANALYTIC BIAS

Many organizations have made efforts to help results from different laboratories agree with one another even if they are not the true result. These efforts, called harmonization, can ensure that results from studies in one laboratory will be reproducible later there or at another institution. Perhaps more importantly, harmonization efforts in the clinical laboratory help patients receive the same standard of care regardless of where they receive their care.

Approaches to harmonizing results vary, and the proper approach depends in part on whether the assays being harmonized are manufactured by in vitro diagnostic companies or have been developed and validated in the laboratory that is conducting the assay. In research, the most common way to harmonize results is to establish a central laboratory as the reference target and to calibrate all assays in the study or consortium to match the central laboratory. The distribution of a research assay to multiple study centers helps bolster the robustness of the study and the potential for translation later by increasing the total number of research participants in the study and by more closely simulating the actual clinical environment in which multiple laboratories will generate clinical results. For the clinical laboratory, organizations such as the WHO have gone to great lengths to develop reference materials that can be used by instrument manufacturers to calibrate their assays to match the consensus mean that is obtained for that reference material when analyzed by many reputable reference laboratories during value assignment. As shown in **Table 1**, reference methods and standards exist for assays that are used in the clinical assessment of iron status and can be used to help harmonize measurements between research studies.

TABLE 1
Reference-method procedures and reference materials available for iron-status assays[1]

	Reference-method procedures[2]	Reference materials	Value assignment[3]
Hemoglobin	Cyanmethemoglobin	WHO 98/708[4]	Consensus[5]
Ferritin	None	WHO 94/572	Consensus
Iron	None	NIST SRM 3126a	ICP-OES[6]
Transferrin	None	IRMM ERM-DA470k/IFCC	Consensus
Soluble transferrin receptor	None	WHO 07/202	Spectrophotometry[7]

[1] ICP-OES, inductively coupled plasma optical emission spectrometry; IFCC, International Federation of Clinical Chemistry and Laboratory Medicine; IRMM, Institute for Reference Materials and Measurements; NIST, National Institute of Standards and Technology.

[2] Listed by the Joint Committee for Traceability in Laboratory Medicine (13).

[3] Methods that are used to establish the concentration of the reference materials. Consensus concentrations are determined as the mean of multiple laboratories and platforms.

[4] Formerly known as CRM 522 (contains hemiglobincyanide).

[5] Many laboratories performed spectrophotometry to determine the concentration.

[6] Although there is no reference-method procedure for iron in serum, the concentration of iron in the pure solution (SRM) was determined via inductively coupled plasma optical emission spectrometry.

[7] Concentration of the pure solution of recombinant protein that was mixed into depleted human serum was determined with absorption at 280 nm.

STANDARDIZATION OF MEASURES: IDENTIFYING AND REMOVING ANALYTIC BIAS

The process of making all assays in all laboratories obtain the same accurate results is called standardization. This lofty goal needs the following 3 pieces of a puzzle that fit together: *1*) a reference-method procedure, *2*) reference materials, and *3*) a site to administer the program of standardization and certification. These resources are used to identify laboratories and assays that are biased compared with the most accurate method available. Assay analytic bias, which can be summarized as the mean ± SD of the bias across many samples compared with the reference-method procedure, can be used to intelligently modify standard operating procedures (e.g., different calibrators or extraction procedure) to generate more accurate results. Reference-method procedures adhere to the guidelines described in International Organization for Standardization (ISO) 15193 (12), which is a consensus document that was established by the ISO (a volunteer-driven nongovernmental organization that provides standards in many different industries) and are approved and cataloged by the Joint Committee for Traceability in Laboratory Medicine (13), which operates under the auspices of the International Bureau of Weights and Standards (an intergovernmental organization attempting to improve quality in analytic sciences). These reference-method procedures are generally laborious and are not fit for general use in clinical laboratories or laboratories that are doing large-scale clinical studies. Along with reference materials that have the amount of analyte carefully determined, e.g., in accordance with the guidelines established in ISO 15194, reference-method procedures will provide the truest result possible. Unfortunately, there is only one reference-method procedure in place for measures of iron status, which is for hemoglobin (13) (Table 1). In addition, as mentioned previously, the reference materials that are available for iron-status bioindicators are mostly assigned via consensus approach, thus making standardization impossible at this time. In any case, when such tools as reference methods and reference materials are available, they can be used by a standardization program to value assign many human samples that can then be used to assess a laboratory's ability to get the correct answers. Manufacturers and laboratories that meet predefined expectations of bias and imprecision may be certified as standardized by the standardization program. Those predefined expectations are based on the clinical use of those assays and are typically developed on the basis of the within-individual variability and between-individual variability of the analyte.

There are several examples of standardization programs that have been developed in nutrition and endocrinology including the National Glycohemoglobin Standardization Program for glycated hemoglobin measurements (14), the Vitamin D Standardization Certification program at the CDC for 25-hydroxyvitamin D (15), and the Hormone Standardization Program at the CDC for testosterone and estradiol (16). Note that the standardization of immunoassays for small molecules can be difficult. The number of molecules that are present in human samples (similar in structure to the analyte of interest or not) that may interfere with the competitive assay approach is staggering. As a result, although standardization can help ensure that, on average, an immunoassay is providing accurate values, there will almost

TABLE 2
Representative proficiency testing results for iron-status assays

Analyte	Most common assay method	Representative all-method imprecision (CV, %)
Hemoglobin	Spectrophotometry	1.8[1]
Iron	Colorimetric dye binding	5.7[2]
Transferrin	Nephelometric immunoassay	4.8[2]
Ferritin	Sandwich immunoassay	16.8[2]

[1] Estimated with an ANOVA with the use of the variability of the reported between-platform means and reported CV of the 5 samples in the Participant Summary Report for 2016-A of the FH1 survey from the College of American Pathologists.

[2] Data are from the Participant Summary Report for 2016-A of the C survey from the College of American Pathologists. Data represent the mean of all-method means for the 5 samples in the survey.

always be sample-specific matrix effects that lead to inaccuracies. Nephelometric immunoassays for high-abundance serum proteins have generally performed better than other immunoassays, but it should be recognized that they detect total protein, including important posttranslational modifications (e.g., a nephelometric transferrin assay detects both glycosylated and unglycosylated transferrin).

PROFICIENCY TESTING

After methods have been developed and sold by instrument manufacturers or developed, validated, and deployed in reference laboratories, the frame of reference for consistency typically shifts to outside of the laboratory. The goal of proficiency testing is to document that assays are behaving as they should during actual use in patient care or clinical research. For many years, proficiency testing programs have used artificial samples or heavily manipulated serum or plasma samples to evaluate performance. This method has been acceptable because assays were compared with themselves and not with each other, such that outlying peers would recognize an issue with their assay and work quickly to solve the problem. Significant discrepancies between platforms were not unexpected or concerning. There are several important organizations that administer proficiency testing programs including the College of American Pathologists (CAP) and the Royal College of Pathologists of Australasia. More recently, as the standard of care begins to rely more and more on guidelines developed in an evidence-based fashion, the clinical interpretation of laboratory tests hinges on cutoffs that are derived from clinical research studies that used assays that were not traceable back to reference materials or reference-method procedures. In an effort to help harmonize those assays that are used in clinical care, particularly when the test interpretation relies on cutoffs, proficiency testing programs are now using minimally processed serum samples as proficiency testing materials. Minimally processed serum samples are drawn from patients, clotted in a controlled fashion, pooled, homogenized, divided into aliquots, and frozen in $<$56 h and are expected to give equivalent results across all platforms (17).

Proficiency testing programs that use this approach include the CAP and the Vitamin D External Quality Assessment Scheme (18). With such testing, as conducted by the CAP, it is possible to evaluate whether methods in production are harmonized. **Table 2** presents the observed variability across all methods for different analytes that are important in the evaluation of iron status. For each of these proficiency testing assessments, several samples were mailed to hundreds of laboratories. The data are used to evaluate each laboratory compared with other laboratories that are running assays of the same analyte. Summary statistics for each analyte, including the CV (calculated as the SD divided by the mean across all measurements), are provided to the participants. From a clinical and analytic chemistry perspective, assays for hemoglobin concentrations collectively perform very well as an indicator of iron-deficiency anemia with a CV of 1.8%. Serum iron and transferrin assays are not quite as good but they are acceptable. However, the serum ferritin (SF) concentration, which is the most commonly used blood-based indicator to identify ID in clinical research studies, has an imprecision of 16.8%. In other words, the 95% CI around a measured concentration of 12 μg/L is 8–16 μg/L. This obviously introduces significant classification errors into clinical research studies. At this time, there is no proficiency testing scheme for soluble transferrin receptor (sTfR) concentrations.

MOVING FORWARD WITH DETERMINING IRON STATUS AND CLINICAL RESEARCH IN IRON METABOLISM

In conclusion, there is currently no consensus around the SF or sTfR concentration cutoffs that should be used to define ID in children or pregnant women. In addition, it is unclear how well assays for SF and sTfR concentrations will agree with one another over time, and there remain many questions surrounding their relevance to health outcomes and the factors that may confound their interpretation. For these assays and for any new assays that will be used to study the importance of iron status in specified outcomes, the field should strive for harmonization. Fortunately, reference materials exist. Funding agencies and journals should immediately require that laboratories analyze these reference materials during the course of their clinical and epidemiologic studies to optimize the traceability of the calibration of the measurement procedures that are used to quantify the concentrations of bioindicators in human samples. In addition, to ensure that results from different studies are comparable over time, federal funding should be set aside to facilitate the accrual of sample sets that can be used in different laboratories around the globe to compare assays with one, which should happen immediately. The results from these sample sets can be used to transform or recalibrate the concentrations in research studies back to a harmonized reference point. Reference materials with more carefully assigned concentrations, which are being explored (J Betz, Office of Dietary Supplements, NIH, personal communication, 2017), will be helpful in the long term, particularly when clinical cutoffs are needed to identify pregnant woman and young children who would benefit from iron supplementation or who are replete and, therefore, not in need of supplementation. It is recommended that the manufacturing of these reference materials also starts immediately.

The author's responsibilities were as follows—the sole author wrote the manuscript and read and approved the final manuscript. The author reported no conflict of interest related to the study.

REFERENCES

1. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature 2012;483:531–3.
2. USPSTF. Iron deficiency anemia in pregnant women: screening and supplementation [Internet]. Rockville (MD): US Preventive Services Task Force; 2016. [cited 2016 Oct 4]. Available from: https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/iron-deficiency-anemia-in-pregnant-women-screening-and-supplementation.
3. USPSTF. Iron deficiency anemia in young children: screening [Internet]. Rockville (MD): US Preventive Services Task Force; 2016. [cited 2016 Oct 4]. Available from: https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/iron-deficiency-anemia-in-young-children-screening?ds=1&s=iron%20deficiency.
4. Pfeiffer CM, Looker AC. Laboratory methodologies for indicators of iron status: strengths, limitations, and analytical challenges. Am J Clin Nutr 2017;106(Suppl):1606S–14S.
5. Ross AC. Impact of chronic and acute inflammation on extra- and intracellular iron homeostasis. Am J Clin Nutr 2017;106(Suppl):1581S–7S.
6. Suchdev PS, Williams AM, Mei Z, Flores-Ayala R, Pasricha SR, Roger LM, Namaste SML. Assessment of iron status in settings of inflammation: challenges and potential approaches. Am J Clin Nutr 2017;106(Suppl):1626S–33S.

7. Westgard JO, Darcy T. The truth about quality: medical usefulness and analytical reliability of laboratory tests. Clin Chim Acta 2004;346:3–11.
8. Laboratory Methods Committee of the Lipid Research Clinics Program of the National Heart Lung and Blood Institute. Cholesterol and triglyceride concentrations in serum/plasma pairs. Clin Chem 1977;23:60–3.
9. Dickerson JA, Laha TJ, Pagano MB, O'Donnell BR, Hoofnagle AN. Improved detection of opioid use in chronic pain patients through monitoring of opioid glucuronides in urine. J Anal Toxicol 2012;36: 541–7.
10. Erali M, Bigelow RB, Meikle AW. ELISA for thyroglobulin in serum: recovery studies to evaluate autoantibody interference and reliability of thyroglobulin values. Clin Chem 1996;42:766–70.
11. Statland BE, Bokelund H, Winkel P. Factors contributing to intra-individual variation of serum constituents: 4. Effects of posture and tourniquet application on variation of serum constituents in healthy subjects. Clin Chem 1974;20:1513–9.
12. International Organization for Standardization (ISO). ISO 15193:2009 in vitro diagnostic medical devices–measurement of quantities in samples of biological origin–requirements for content and presentation of reference measurement procedures [Internet]. Geneva (Switzerland): International Organization for Standardization; 2009. [cited 2017 Feb 13]. Available from: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=42021.
13. Joint Committee for Traceability in Laboratory Medicine. JCTLM database: laboratory medicine and in vitro diagnostics [Internet]. Sèvres (France): BIPM; 2017. [cited 2017 Feb 10]. Available from: http://www.bipm.org/jctlm/.
14. Little RR, Rohlfing CL, Sacks DB; National Glycohemoglobin Standardization Program Steering Committee. Status of hemoglobin A1c measurement and goals for improvement: from chaos to order for improving diabetes care. Clin Chem 2011;57:205–14.
15. Binkley N, Sempos CT, Vitamin DSP. Standardizing vitamin D assays: the way forward. J Bone Miner Res 2014;29:1709–14.
16. Yun YM, Botelho JC, Chandler DW, Katayev A, Roberts WL, Stanczyk FZ, Vesper HW, Nakamoto JM, Garibaldi L, Clarke NJ, et al. Performance criteria for testosterone measurements based on biological variation in adult males: recommendations from the Partnership for the Accurate Testing of Hormones. Clin Chem 2012;58: 1703–10.
17. Stöckl D, Libeer JC, Reinauer H, Thienpont LM, De Leenheer AP. Accuracy-based assessment of proficiency testing results with serum from single donations: possibilities and limitations. Clin Chem 1996; 42:469–70.
18. DEQAS Advisory Panel. Vitamin D External Quality Assessment Scheme (DEQAS) [Internet]. London: DEQAS; 2017. [cited 2017 Feb 13]. Available from: http://www.deqas.org/.