

The Biological Variation Data Critical Appraisal Checklist: A Standard for Evaluating Studies on Biological Variation

Aasne K. Aarsand,^{1,2*} Thomas Røraas,² Pilar Fernandez-Calle,^{3,4} Carmen Ricos,⁴ Jorge Díaz-Garzón,^{3,4} Niels Jonker,⁵ Carmen Perich,^{4,6} Elisabet González-Lao,^{4,7} Anna Carobene,⁸ Joana Minchinela,^{4,9} Abdurrahman Coşkun,¹⁰ Margarita Simón,^{4,11} Virtudes Álvarez,⁴ William A. Bartlett,¹² Pilar Fernández-Fernández,⁴ Beatriz Boned,^{4,13} Federica Braga,¹⁴ Zoraida Corte,^{4,15} Berna Aslan,¹⁶ and Sverre Sandberg^{1,2,17} on behalf of the European Federation of Clinical Chemistry and Laboratory Medicine Working Group on Biological Variation and Task and Finish Group for the Biological Variation Database

BACKGROUND: Concern has been raised about the quality of available biological variation (BV) estimates and the effect of their application in clinical practice. A European Federation of Clinical Chemistry and Laboratory Medicine Task and Finish Group has addressed this issue. The aim of this report is to (a) describe the Biological Variation Data Critical Appraisal Checklist (BIVAC), which verifies whether publications have included all essential elements that may impact the veracity of associated BV estimates, (b) use the BIVAC to critically appraise existing BV publications on enzymes, lipids, kidney, and diabetes-related measurands, and (c) apply metaanalysis to deliver a global within-subject BV (CV_I) estimate for alanine aminotransferase (ALT).

METHODS: In the BIVAC, publications were rated as A, B, C, or D, indicating descending compliance for 14 BIVAC quality items, focusing on study design, methodology, and statistical handling. A D grade indicated that associated BV estimates should not be applied in clinical practice. Systematic searches were applied to identify BV studies for 28 different measurands.

RESULTS: In total, 128 publications were identified, providing 935 different BV estimates. Nine percent achieved

D scores. Outlier analysis and variance homogeneity testing were scored as C in >60% of 847 cases. Metaanalysis delivered a CV_I estimate for ALT of 15.4%.

CONCLUSIONS: Application of BIVAC to BV publications identified deficiencies in required study detail and delivery, especially for statistical analysis. Those deficiencies impact the veracity of BV estimates. BV data from BIVAC-compliant studies can be combined to deliver robust global estimates for safe clinical application.

© 2017 American Association for Clinical Chemistry

Biological variation (BV)¹⁸ data have many applications, most importantly being used to aid in diagnosing and monitoring disease and for setting analytical performance specifications (1). Safe clinical application of BV data requires estimates to be reliable and representative of the specific population group and situation to which they are applied. The current main collated source of BV data is the online 2014 BV Database hosted on the Westgard website (2, 3). Here, within-subject (CV_I) and between-subject (CV_G) BV estimates of a range of measurands with their associated analytical performance specifica-

¹ Norwegian Porphyria Centre, Laboratory of Clinical Biochemistry, Haukeland University Hospital, Bergen, Norway; ² Norwegian Quality Improvement of Laboratory Examinations (NOKLUS), Haraldsplass Deaconess Hospital, Bergen, Norway; ³ La Paz University Hospital, Madrid, Spain; ⁴ Spanish Society of Laboratory Medicine (SEQC-ML), Analytical Quality Commission, Barcelona, Spain; ⁵ Certe, Wilhelmina Ziekenhuis Assen, Assen, the Netherlands; ⁶ Clinic Laboratory Hospital Vall d'Hebron, Barcelona, Spain; ⁷ Catlab, Clinic Laboratory, Mutua Terrassa University Hospital, Barcelona, Spain; ⁸ Servizio Medicina di Laboratorio, Ospedale San Raffaele, Milan, Italy; ⁹ Metropolitana Nord Unified Laboratory (LUMN), Germans Trias i Pujol University Hospital, Badalona, Spain; ¹⁰ Acibadem University, School of Medicine, Atasehir, Istanbul, Turkey; ¹¹ Laboratori de l'Alt Penedès, l'Anoia i el Garraf, Barcelona, Spain; ¹² Blood Sciences, Ninewells Hospital and Medical School, Scotland, UK; ¹³ Royo Villanova Hospital, Zaragoza, Spain; ¹⁴ Research Centre for Metrological Traceability in Laboratory Medicine (CIRME), University of Milan, Milan, Italy; ¹⁵ San Agustín University Hospital, Aviles, Asturias, Spain; ¹⁶ Institute for Quality Management in Healthcare (IQMH), Centre for Proficiency Testing, Toronto, ON,

Canada; ¹⁷ Department of Global Health and Primary Care, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway.

* Address correspondence to this author at: Norwegian Porphyria Centre, Laboratory of Clinical Biochemistry, Haukeland University Hospital, NO-5021 Bergen, Norway. Fax +47-55-97-31-15; e-mail aasne.aarsand@helse-bergen.no.

Received September 14, 2017; accepted November 16, 2017.

Previously published online at DOI: 10.1373/clinchem.2017.281808

© 2017 American Association for Clinical Chemistry

¹⁸ Nonstandard abbreviations: BV, biological variation; CV_I , within-subject biological variation; CV_G , between-subject biological variation; EFLM, European Federation of Clinical Chemistry and Laboratory Medicine; TFG-BVD, Task and Finish Group for the Biological Variation Database; BIVAC, Biological Variation Data Critical Appraisal Checklist; ALT, alanine aminotransferase; WG-BV, Working Group on Biological Variation; QI, quality item; CV_A , analytical CV; HbA1c, hemoglobin A_{1c}; EuBIVAS, European Biological Variation Study.

tions are presented. The data in the database are derived from available BV studies and have been updated every 2 years, up until 2014 (4). This database has delivered a useful source of BV data, but questions have been raised regarding the veracity of the estimates presented and the potential impact this may have when applied to clinical practice (5–7). Following the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) in 2014, an EFLM Task and Finish Group for the BV Database (TFG-BVD) was established (8, 9). Among its objectives was to develop a critical appraisal list for evaluation of BV literature. This report aims to (a) describe the development and content of this checklist, the Biological Variation Data Critical Appraisal Checklist (BIVAC), (b) present the results of applying the BIVAC to BV studies on enzymes, lipids, kidney, and diabetes-related measurands, and (c) as an example, deliver a global estimate for CV_1 for serum alanine aminotransferase (ALT) based on metaanalysis of critically appraised BV studies.

Material and Methods

DEVELOPMENT OF THE BIVAC

The TFG-BVD comprised members from the EFLM Working Group on BV (WG-BV) (10), the Analytical Quality Commission of the Spanish Society of Laboratory Medicine, and others with extensive experience in theoretical and practical aspects of generating data on BV. The checklist for reporting of BV data, recently published by the EFLM WG-BV (11), was used as a starting point to identify essential elements that provided the basis for a draft checklist. To assess whether the proposed checklist captured all essential elements, as well as to ascertain whether the scoring of articles using this checklist was consistent, it was applied individually by 21 people on the same 25 randomly selected BV publications. This enabled fine-tuning of the quality items (QIs) and associated scores. Thereafter, experts in the TFG-BVD, who had not participated in the practical scoring process, reviewed the draft checklist. This approach of scoring BV publications followed by revision of the checklist was repeated 3 times, resulting in a final consensus-based checklist.

DESCRIPTION OF THE BIVAC

The BIVAC focuses primarily on effects of study design, the applied measurement procedure, and statistical handling of data on CV_1 estimates. When applying the BIVAC, the publication is rated with regard to 14 QIs scored as A, B, C, or D (Table 1). An overall grade A indicates full compliance with all the 14 BIVAC QIs; a grade B is applied if the lowest QI score achieved is a B. Similarly, the publication is graded C if the lowest QI score is C. D is a scoring option for QIs that are consid-

ered essential for measures of BV to be reliable. In the BIVAC scoring system, the QIs associated with the grade are given as a subscript; a grade of $C_{5,10}$ indicates that QI numbers 5 and 10 were scored as C.

CRITICAL APPRAISAL OF BV STUDIES

Systematic searches for BV studies were performed by 4 different groups of assessors for enzymes, lipids, kidney, and diabetes-related measurands (Table 2). Relevant studies were identified by first including all articles for these measurands listed in the online 2014 BV Database (2). Additionally, searches were carried out in PubMed, with the cut-date of December 31, 2016, using as key terms [the measurand] in question with each of the following combinations: “within-subject *,” “between-subject *,” “within-person *,” “between-person *,” “interindividual *,” and “intraindividual *,” where the asterisk denotes “biological variation,” “variation,” “coefficient of variation,” and “CV” (12). Reference lists in the retrieved papers were also checked for other relevant publications. Articles published in languages other than English, Spanish, or Italian and those for which full-text versions were irretrievable, were excluded.

Sets of 2 assessors independently scored all the publications containing data on BV. When several BV estimates were reported in the same publication, either for different measurands or 1 measurand in different settings (e.g., healthy subjects/diseased, different sex/age groups, short-term/long-term data), scoring and data extraction were performed for each of the individually reported estimates. If the results for any of the QIs were not in accordance with each other, a third assessor performed a completely new review, followed, if necessary, by discussions between the assessors or the full group to achieve agreement. To summarize scoring results per publication, the BIVAC QI scores for 1 arbitrarily selected measurand from each publication were used.

CALCULATION OF A GLOBAL CV_1 ESTIMATE

To calculate a global CV_1 estimate, a metaanalysis approach may be used. When performing metaanalysis, the inverse of each of the collected studies’ variances is commonly used as weight (13). For an estimate of CV_1 , the best measure of variability is the CI, which combines information from both the analytical CV (CV_A) and the number of subjects, samples, and replicates. CIs for CV_1 were calculated as previously described (14) for those ALT studies that reported the required information. In our metaanalysis, the inverse of the width of the CI was used as weight. To reflect the lack of data treatment for potential confounders, such as outliers or systematic effects that could inflate the estimated CV_1 , the different quality grades were arbitrarily given weights: A = 4, B = 2, and C = 1. To provide the global CV_1 estimate, the weighted median was used (15), available in the

Table 1. BIVAC with criteria for achieving A, B, C, and D scores for the different quality items and their rationale.

QI	Quality question	Quality scoring				Rationale
		A	B	C	D ^a	
Material and methods						
1: Scale	Is the measurand given on a ratio scale?	Yes	No	–	–	Only the ratio scale has a meaningful zero, and estimation of within-subject biological variation (CV _I) for measurands on nonratio scales requires special attention.
2: Subjects	Are subjects and the population documented in terms of (a) Number of participants (b) Sex (c) Age/age-group (d) State of well-being/health status?	Yes	All criteria fulfilled, but participants are described as healthy volunteers without any further details given. (a) and (d) are documented; data on (b) gender and/or (c) age/age-group is lacking, but this is not of importance for the measurand in question.	(a) and (d) are documented, but data on (b) sex and/or (c) age/age group is lacking, and this is of importance for the measurand in question.	There is no information on (a) number of participants or (d) state of well-being/health status.	Necessary to characterize the population in which the study has been performed and thereby to allow for comparisons with other studies.

Continued on page 504

Continued on page 504

Table 1. BIVAC with criteria for achieving A, B, C, and D scores for the different quality items and their rationale. (Continued from page 503)

QI	Quality question	Quality scoring				Rationale
		A	B	C	D ^a	
3: Samples	Are the following documented? (a) Number of samples collected (b) Type of sample material used (c) Timing of sample collections (d) Length of study period	Yes	(a), (c), and (d) are documented. (b) Sample material; insufficient detail is given, but this is not of importance for the measurand in question.	(a), (c), and (d) are documented, but no information or insufficient detail on (b) sample material is given, with consequences for the measurand in question.	(a) Number of samples, (c) timing of sample collections, and/or (d) length of study period are not presented or deducible.	Necessary to characterize how and in what material the study has been performed and to evaluate if variation in timing gives different estimates.
4: Measurand	Are the measurand and the measurement procedure documented?	Yes A detailed method description is presented Or A reference to article where the method is described in detail is provided, Or An identifiable method has been applied and is described with sufficient detail. ^b	Insufficient detail on the method is given, but it is not of importance for the measurand in question.	Insufficient detail on the method is given or an outdated method has been used, which may be of importance for the measurand in question.	The method is obsolete and no longer valid, i.e., methods in use today estimate another measurand, or the method is not fit for the purpose of estimating biological variation in the chosen population.	Adequate description of the measurand and measurement procedure is necessary to ensure transferability of data.

Continued on page 505

Table 1. BIVAC with criteria for achieving A, B, C, and D scores for the different quality items and their rationale. (Continued from page 504)

QI	Quality question	Quality scoring				Rationale
		A	B	C	D ^a	
5: Preanalytical procedures	Are preanalytical procedures described and standardized to minimize preanalytical variation?	Yes	Insufficient detail on preanalytical treatment is given, but it is unlikely to be of importance for the measurand in question.	Insufficient detail on preanalytical treatment is given, which may be of importance for the measurand in question. No details on preanalytical procedures/ treatment given.	–	Appropriate preanalytical procedures are necessary to avoid that preanalytical variation affects the CV _I estimate.
6: Estimates of analytical variation	Are estimates of analytical variation based on replicate analysis, and are estimates presented?	Yes, estimates are presented, with all replicates for the same subject having been analyzed in the same run.	Estimates are presented but have been obtained by other method than replicate analysis or replicate analyses of samples have been performed in different runs.	No estimates are presented.	–	Replicate analysis performed in the same run provides the most correct estimate of analytical variation when calculating the CV _I .
7: Steady state	Are all included individuals in steady state, or have data been adequately transformed?	Yes	Individual trend analysis has not been performed, but this is unlikely to be of importance for the measurand in question.	Individual trend analysis has not been performed, and this may be of importance for the measurand (e.g., hormones) or clinical setting (e.g., diseased subjects).	–	For the obtained estimate of CV _I to be reliable, subjects must be in steady state.

Continued on page 506

Table 1. BIVAC with criteria for achieving A, B, C, and D scores for the different quality items and their rationale. (Continued from page 505)

QI	Quality question	Quality scoring				Rationale
		A	B	C	D ^a	
8: Outliers	Has testing for outliers of (a) Replicates (b) Samples per subject (c) Subjects been performed?	Yes	(b) Fulfilled, but Replicate analysis of samples has not been performed, and/or Testing for outliers of replicates has not been performed. and/or Outlier analysis for subjects has not been performed.	Outlier analysis of (b) samples per subjects has not been performed. Article states that outlier analysis has been performed without giving any further details. Outlier analysis has been performed on only the total data set for all participants.	—	Lack of removal of outliers between duplicates may falsely increase or decrease the estimated CV _i . Outliers within a subject's series usually increase the estimated CV _i and might also cause heterogeneity.
9: Normally distributed data	Has the distribution of data for each subject been assessed for normality and, if not conforming to a normal distribution, been appropriately transformed?	Yes	No	—	—	Normality is not a prerequisite for estimation of SD _i , but distribution is relevant for estimation of CV _i . CV-ANOVA can be applied independently of the distribution. The normality distribution is necessary for CI and reference change values (RCV) to be calculated directly.
10: Variance homogeneity	Has variance homogeneity been examined? ^c	Yes	—	No	—	Heterogeneity of variances will cause estimates not to be generalizable to the whole population.

Continued on page 507

Table 1. BIVAC with criteria for achieving A, B, C, and D scores for the different quality items and their rationale. (Continued from page 506)

QI	Quality question	Quality scoring				Rationale
		A	B	C	D ^a	
11: Statistical method	Is the statistical method applied a (nested) ANOVA or an equivalent variance decomposition model ^d with estimation of analytical, within-, and (if relevant) between-subject variation directly?	Yes	Simple subtraction of variances used.	Other method or method not declared.	–	The nested ANOVA (or equivalent variance decomposition model) provides the most accurate estimate of the individual variances, especially for nonbalanced data.
Results						
12: Confidence limits	Are confidence limits around estimates of CV _i presented or are they possible to calculate? ^e	Yes	–	No	–	Calculations of confidence limits are necessary for evaluation of the CV _i estimate and when comparing estimates obtained in different studies and/or clinical situations.

Continued on page 508

Table 1. BIVAC with criteria for achieving A, B, C, and D scores for the different quality items and their rationale. (Continued from page 507)

QI	Quality question	Quality scoring				Rationale
		A	B	C	D ^a	
13: Number of included results	Is the number of results excluded following analysis of outliers and homogeneity of variances given, or can they be deduced from the presented data?	Yes	Outlier and variance homogeneity analyses have been performed, but the number of excluded results is not presented, and only the total number of results used for estimation of BV is documented.	(a) The number of results used for estimation of BV is not documented. (b) Analysis of outliers and/or homogeneity of variances have not been performed.	—	Presentation of the number of data points used to estimate CV _i is important to address group homogeneity and the quality of sample collection.
14: Concentrations	Are mean concentrations of the measurand(s) presented or can they be extracted from figures?	Yes	No	—	—	Necessary for the evaluation of the correlation between CV _i and concentrations.

^a Quality items that can receive a D score are regarded as critical for data quality, and estimates from a study receiving any D grade(s) are considered unsuitable for application in clinical practise.

^b Detailed description includes which method has been applied, on which instrument the method has been run, and the name of producer of the instrument.

^c Acceptable tests for assessing variance homogeneity include Bartlett, Cochran, Levene, Brown-Forsythe, and Fligner-Killeen.

^d The applied method must be a multilevel/hierarchical/nested random- or mixed-effects model. Appropriate models are ANOVA, analysis of covariance (ANCOVA), linear mixed models (LMM), general linear models (GLM), general linear mixed models (GLMM), generalized linear models, and generalized estimating equations (GEE). A mixed-effects model is characterized by both random (inter, intra, analytical, . . .) and fixed effect (sex, age, . . .) being present.

^e Calculation of confidence limits requires data on mean number of samples and of subjects used in the calculation of the BV estimates and estimates of analytical variation.

Table 2. Overview of measurands included in the systematic review.^a

Lipids	Enzymes	Diabetes-related measurands	Kidney-related measurands
Apolipoprotein A1	Alanine aminotransferase	Adiponectin	Albumin
Apolipoprotein B	Aspartate aminotransferase	C-peptide	Urine albumin
HDL cholesterol	γ -Glutamyl transferase	Blood hemoglobin A1c	Creatinine
LDL cholesterol	Lactate dehydrogenase	Fructosamine	Chloride
Estimated LDL cholesterol		Glucose	Cystatin C
Total cholesterol		Insulin	Potassium
Triglycerides		Insulin-like growth factor 1	Sodium
		Insulin-like growth factor-binding protein 3	Urea
		Lactate	
		Pyruvate	

^a All measurands are in serum and/or plasma unless otherwise specified.

R package matrixStats (16), for which the combined result of the inverse width of the CI and the quality score decided the weight of each estimate. Only estimates from healthy adults for whom sampling was performed weekly were included in the metaanalysis. When publications reported separate estimates for males and females, these estimates were combined to provide a common estimate by applying the weighted mean on the point estimates and corresponding CIs. For the global CV_I, a bias-corrected bootstrap approach (17) was used to indicate measures of uncertainty, which for few estimates correspond to the range.

Results

REVIEW AND APPRAISAL OF BV PUBLICATIONS

The literature review identified 128 publications reporting BV data for the selected measurands (see File 1 in the

Data Supplement that accompanies the online version of this article at <http://www.clinchem.org/content/vol64/issue3>), after exclusion of 5 full-text articles that were irretrievable and 2 articles that were duplicate publications in Spanish and English. The online 2014 BV Database was the best source for identifying relevant publications, as exemplified for creatinine (Table 3). Most publications included BV estimates for several measurands, demographic groups, clinical settings, and/or sampling intervals. In total, 117 publications were scored as A, B, or C for at least 1 measurand, delivering 847 different BV estimates. Each of QIs outliers (QI 8), variance homogeneity (QI 10), and number of included results (QI 13) received score C in >50% of the articles (Table 4). Thirty-seven publications received a D score for at least 1 measurand (26 of the 117 and an additional 11 for which a D score was awarded for all included measurands), delivering 88 different BV estimates, mostly for enzymes (73%). Eighty percent

Table 3. Results for different PubMed search strategies for publications on BV of creatinine and the number of articles considered relevant after review.^a

Keywords	Targeted fields	Number of hits	Number of relevant articles
Variation and creatinine	All	2343	Not assessed
Biological variation and creatinine	All	543	13
Biological variation and creatinine	Title	5	4
Biological and creatinine	Title	38	4
Variation and creatinine	Title	49	7
Search strategy as described in Methods	Title, abstract	402	19
Online 2014 BV Database			40

^a The online 2014 BV Database contains 40 publications with BV data for creatinine, 8 of which are not included in PubMed.

Table 4. Results for Biological Variation Data Critical Appraisal Checklist A, B, and C scores for (a) 117 articles containing BV estimates for lipids, enzymes, diabetes, and/or kidney-related measurands (scores for 1 arbitrarily selected measurand from each publication) and (b) 847 individual BV estimates reported in these publications.^a

Quality item number	Quality item	A% (n = 117 articles)	A% (n = 847 estimates)	B% (n = 117 articles)	B% (n = 847 estimates)	C% (n = 117 articles)	C% (n = 847 estimates)
1	Scale	94.0	97.8	6.0	2.2	—	—
2	Subjects	88.0	94.5	6.8	3.8	5.1	1.8
3	Samples	88.9	93.7	8.5	5.4	2.6	0.8
4	Measurand	89.7	91.4	6.8	4.3	3.4	4.4
5	Preamalytical procedures	80.3	88.7	13.7	7.3	6.0	4.0
6	Estimates of analytical variation	50.4	48.4	40.2	48.4	9.4	3.2
7	Steady state	28.2	39.1	64.1	56.2	7.7	4.7
8	Outliers	22.2	16.2	11.1	21.4	66.7	62.5
9	Normally distributed data	8.5	15.3	91.5	84.7	—	—
10	Variance homogeneity	17.9	25.4	—	—	82.1	74.6
11	Statistical method	45.3	45.3	41.0	46.2	13.7	8.5
12	Confidence limits	76.1	81.9	—	—	23.9	18.1
13	Number of included results	32.5	30.0	6.8	16.4	60.7	53.6
14	Concentrations	82.9	90.2	17.1	9.8	—	—

^a Articles receiving D scores for all included measurands (n = 11) were excluded.

of all D scores were awarded for QI 4 (measurand). On average, for the whole review period, the assessors disagreed on ≥ 1 QIs for about 30% of the included papers. The most frequent cause of disagreement was for enzymes—the QI 4 (measurand)—and for lipids, kidney, and diabetes-related measurand QIs 8-Outliers, 9-Normally distributed data, 10-Variance homogeneity, and, thus indirectly, 13-Number of included results. For diabetes-related measurands, discussions were also required to determine which preanalytical requirements were necessary for an A score.

METANALYSIS OF CV_I ESTIMATES FOR SERUM ALT

Twenty-three studies were identified for ALT, with 1 article receiving an overall B grade, 6 C, and 16 D (Table 5; see also File 2 in the online Data Supplement), with D articles being excluded from further analysis. Four additional studies were excluded because 1 was performed in children (18), sampling was not weekly in 2 (19, 20), and sampling was biweekly and data required for calculation of CIs were lacking in 1 (21). Two reported identical results from the same healthy control group (22, 23); thus, only 2 C studies performed in healthy adults with weekly sampling were included in the meta-analysis (23, 24). Estimates were reported separately for males and females in both (Table 5), resulting in common CV_I estimates at 15.2% (CI, 13.7–17.0) (23) and 15.7% (CI, 12.4–17.2) (24), respectively. The combi-

nation of these delivered a global CV_I estimate of 15.4% (range, 15.2–15.7).

Discussion

Biological variation data are essential in everyday laboratory work, delivering a requirement that they must be robust and reliable. However, currently, estimates provided for the same measurand may vary substantially. This may have considerable impact when applying BV estimates to set analytical performance specifications and when used in applications aimed at enabling accurate diagnosis and monitoring of patients. In existing studies, there is variation in protocols and methodological approaches that may impact the veracity of the published BV estimates. The BIVAC provides a tool to assess the quality of BV publications by verifying whether all essential elements that may impact on veracity and utility of the data are present. This is important in the context of assessment of the literature on BV that stretches back >40 years and in driving up the quality of future studies of BV. This approach also enables estimates to be pooled from compatible studies to provide more robust global BV estimates.

LITERATURE REVIEW

It may be challenging to identify relevant BV publications because there has not been a Medical Subject Heading specifically for BV; authors use different terms to

Table 5. Scoring of BV studies on serum alanine aminotransferase by the BV Data Critical Appraisal Checklist.^a

Author	Year	Reference	QI														Summary score	Sampling intervals	CV _i (95% CI)
			1	2	3	4	5	6	7	8	9	10	11	12	13	14			
Statland	1973	19	A	A	A	C	A	A	B	C	B	C	B	A	C	A	C _{4,8,10,13}	3/day	6.7 (4.7–9.9)
Winkel	1974	21	A	A	A	C	A	A	B	C	B	C	A	C	C	A	C _{4,8,10,12,13}	2/week	26.4
Hölzel	1987	22	A	A	A	C	B	A	B	C	A	A	B	A	C	A	C _{4,8,13}	1/week	F: 15.7 (13.7–18.4) M: 14.4 (12.2–17.4)
Hölzel	1987	23	A	A	A	C	B	A	B	C	A	A	B	A	C	A	C _{4,8,13}	1/week	F: 15.7 (13.7–18.4) M: 14.4 (12.2–17.4)
Pineda-Tenor	2013	24	A	A	A	A	B	B	B	C	B	A	B	A	C	A	C _{8,13}	1/week	F: 17.3 (15.3–19.5) M: 14 (12–16)
Bailey	2014	18	A	A	A	A	A	B	B	B	B	A	B	A	B	A	B _{6,7,8,9,11,13}	4/day	15.6 (11.8–18.9)
Qi	2016	20	A	A	A	A	A	A	B	C	A	C	A	A	A	A	C _{8,10}	1/day	3.57 (3.2–4.1)

^a Publications receiving quality score D for any item are not included (n = 16; see File 2 in the online Data Supplement). Listed estimates of within-subject BV (CV_i) are for healthy adults if provided, separated into results for males (M) and females (F) when reported.

describe the BV components (12), and hits are likely to include many irrelevant publications, as exemplified for creatinine (Table 3). The online 2014 BV Database is the result of many years of work reviewing literature on BV, and it delivered the majority of publications included in our review. To improve future classification of BV publications, the WG-BV has suggested and has had accepted a Medical Subject Heading term for BV, expected to be available from December 2017.

BIVAC AND ITS APPLICATION

To reflect the importance for the reliability of the resulting BV estimates, the QIs included in the BIVAC have been given different importance. Some are identified as critical, in that failure to meet the QI will render estimates unsuitable for use. The BIVAC identifies studies not meeting these as D studies, which is the lowest of the A to D ratings. Other QIs that do not directly impact the BV estimates themselves may have consequences for interpretation of the data and their application. The rationale for each QI in the checklist follows below. It is important to be aware that the BIVAC, similar to GRADE for guidelines (25), mainly addresses the methodology used and reported. Thus, even if an article does not specifically address an item in the BIVAC, e.g., outlier analysis, it cannot be ruled out that this has indeed been assessed and provided for, but that the authors have failed to include this in their method description. The BIVAC scoring system grade reflects the lowest score given to any of the 14 QIs, thus providing a transparent review of where there may be concerns with the study. Other scoring systems such as a grade point average may also have been appropriate and were discussed. However, the BIVAC scoring system was chosen because it makes

immediately apparent which critical elements are lacking, thus providing the reader the opportunity to take this into consideration when reviewing and applying the data.

The BIVAC has been developed as a checklist for the detailed review of BV studies. Disagreements between assessors generally concerned the 3 statistical QIs 8, 9, and 10, for which the differing scores mostly had their basis in the lack of details or unclear description provided in the assessed publications. Some of these articles were reviewed by ≥ 2 groups because they contained results for many different measurands. A comparison between the different groups' scores for these QIs was also performed. This revealed some inconsistencies in the interpretation of the statistical elements in these mostly older publications. A harmonized score for these publications was agreed on in the larger group, whereupon scoring results were updated accordingly and other publications were reassessed if relevant. To ensure the harmonized scoring of these and other QIs that caused discussions, the BIVAC checklist was expanded with subscripts with required details on the analytical method, acceptable tests for assessing variance homogeneity, and appropriate statistical methods (Table 1). This subsequently led to less disagreement between assessors during the later part of the review process. Additionally, some of the QIs, such as QI 4 Measurand and QI 5 Preanalytical procedures, include a subjective evaluation as to which detail is required for the different scores. It would be preferable to have a standardized approach, but this will vary from measurand to measurand. Therefore, the requirements for the different scores for QI 4 and QI 5 in our study are based on the opinion of the expert group, which may be considered a limitation of the BIVAC. Successful appli-

cation of the BIVAC checklist depends on knowledge of the measurand that is being appraised and of the statistics related to estimating components of BV. However, most importantly, it depends on the details provided in the publications on how the study has been performed and how the BV estimates have been obtained.

BIVAC QUALITY ITEMS

QI 1: Scale. This QI reviews whether the measurand is given on a ratio scale, which is important because only the ratio scale has a meaningful zero. Thus, estimation of CV_I for measurands on nonratio scales requires special attention (26, 27). This can be exemplified by hemoglobin A_{1c} (HbA_{1c}), which can be expressed both in IFCC (mmol/mol) and Diabetes Control and Complications Trial (percentage) units. HbA_{1c} in Diabetes Control and Complications Trial units has no true zero value and, thus, is estimated on the interval scale, whereas HbA_{1c} in IFCC units is measured on a ratio scale. Seven of the reviewed publications included measurands assessed on a nonratio scale, i.e., HbA_{1c} given as percentage.

QIs 2, 3, and 4: Subjects, samples, and the measurand. The BIVAC QIs relating to subjects (QI 2) and samples (QI 3) are considered critical for the reliability of the associated measures of BV. BV data are reference data, and their safe clinical application across populations necessitates that the attributes of the population from which they were derived are adequately characterized and reported. Furthermore, details on number of subjects, number of samples, and sample material are necessary to compare with other studies and to allow for the generation of global BV estimates. To evaluate whether different sampling intervals lead to different CV_I estimates, information on timing of samples is required. As expected, most publications provided adequate details on this. The measurement procedure, QI 4, is also essential because older generation analytical methods may deliver estimates for a different measurand. The majorities of D scores were for publications on enzymes, with obsolete methods being the main reason for this classification. This is not surprising given that there have been substantial changes in the analytical principles for enzymes in the past decades, and most of these articles predated the current IFCC recommendations for enzyme measurement optimization (28).

QIs 5, 6, and 7: Preanalytical procedures, estimates of analytical variation, and steady state. Standardized and appropriate preanalytical procedures are necessary for obtaining trustworthy BV estimates. If this requirement is not fulfilled, increased preanalytical variation may transfer into an overestimation of CV_I (and CV_G). Most of the publications assessed in our study were considered to have applied adequate preanalytical procedures (Table

4). Obtaining accurate CV_A estimates (QI 6) is also important for providing reliable BV estimates and CIs (29). A suboptimal approach for establishing the CV_A is the use of internal quality control data generated based on commercial sample materials or patient samples. **The recommended approach is by replicate analysis of the study samples, preferably by analyzing all samples from the same subject in duplicate within the same series.** In half of the articles included in our study, this method was used to estimate the CV_A . **Additionally, there should be no systematic change in the concentration of the measurand during the study period** (QI 7 Steady state), or if there is, data must be adequately transformed. For many measurands examined in healthy subjects, changes in concentrations are unlikely, but for some measurands, such as hormones, this can be expected. **In studies of nonhealthy subjects, care must be taken to ensure a steady-state situation** (30). In 28% of the assessed publications, some type of trend analysis had been performed, or it was considered that data had been adequately transformed.

QIs 8, 9, and 10: Outliers, normality, and variance homogeneity. Concern has been raised regarding the quality of the statistical approach in many BV publications (5, 6), such as analysis for outliers (QI 8) and variance homogeneity (QI 10). In our study, these 2 QIs were scored as C for the majority of publications (Table 4). Failure to identify and consequently remove outliers between duplicates may result in both overestimation and underestimation of the CV_I , whereas failure to remove outliers within a subject's series may lead to an overestimation of the CV_I . The omission to identify and take action if there is variance heterogeneity will cause estimates not to be generalizable to a general setting. Our study clearly identifies these 2 statistical elements as potential contributors to the varying BV estimates published for the same measurand. QI 9 assesses whether the distribution of data for each person has been assessed for normality and, if not confirming to this, has been appropriately transformed. The distribution is of less importance when applying a CV-ANOVA (31), but normality is important if outlier and homogeneity testing or CIs depend on the normality assumption, which they typically do. Furthermore, it is a prerequisite for the direct calculation of reference change values (31, 32). It is important to be aware that the normality item is not related to the distribution of the results from the whole data set but rather the distribution of each of the different model effects: analytical, within-subject, and, to a lesser degree, between-subject (31). Testing of normality for the analytical effect can be done on the pooled standardized residuals (33). For the within-subject effect, one can either test the normality of each subject or again use the pooled standardized residuals. In <10% of publications, it was considered that normality testing or appropriate transformation had been performed.

QI 11: Statistical method. Wide-ranging statistical methods are applied for the estimation of BV data. A multilevel/hierarchical/nested random- or mixed-effects model delivers the most accurate estimates by providing results for CV_I , CV_G , and CV_A directly, with CV-ANOVA being a simple and robust method (31). In our study, 45% had applied a variance decomposition model.

QI 12: Confidence intervals. Until recently, it has been uncommon to accompany BV estimates with measures of uncertainty (QI 12), as exemplified in ALT, for which only 1 (20) of 7 non-D studies provided this. Confidence limits for CV_I can, however, be calculated when appropriate CV_A estimates, number of samples, and number of subjects used in the BV estimates are provided in the publication (14). In the BIVAC, the omission of reporting CIs will not generate a C score itself, but the omission of the necessary data elements for their calculation will. These basic elements were missing in 18% of scores.

QIs 13 and 14: Number of included results and concentrations. Only 30% of publications stated the number of results (QI 13) that were used as the basis for the BV estimates, i.e., after exclusion of data following outlier and variance homogeneity testing. This is important for addressing group homogeneity and the generalizability of data. In most publications, mean concentrations of the measurands were presented or could be extracted from the results (QI 14). This is necessary to evaluate the correlation between CV_I and concentration; however, it does not affect the reliability of the BV estimates themselves.

METAAANALYSIS OF CV_I ESTIMATES

A metaanalysis of ALT articles in which estimates were based on weekly sampling in healthy adults provided a global CV_I estimate of 15.4% (range, 15.2–15.7). However, of the 23 published articles on ALT, only 2 relevant studies could be included. In the online 2014 BV Database, the CV_I for ALT is given as 19.4%; that is the median CV_I based on 9 studies (2), most of which do not meet the BIVAC criteria. Updated BV estimates for enzymes from the EFLM European Biological Variation Study (EuBIVAS) have recently been published (34). EuBIVAS used contemporary analytical methods and applied a stringent preanalytical, analytical, and statistical protocol with samples from 91 healthy individuals. When appraising this publication by the BIVAC, it received an overall score of A and reported a CV_I for ALT of 9.3% (95% CI, 8.7–10.0). This is clearly lower than both the estimate given in the online 2014 BV Database and the estimate provided by metaanalysis in our study. However, when the EuBIVAS CV_I estimate was included in our metaanalysis, where it was given high weight because of both the narrow CI and the quality score A, an

updated metaanalysis estimate of 10.2% (range, 9.3–15.7) was obtained. The online 2014 BV Database and EuBIVAS/updated metaanalysis estimate would deliver very different performance specifications for the assay of ALT. This underlines the need for high quality studies and the application of appropriate methods to deliver reliable global estimates. When grade A studies are available, using estimates from such studies alone may be considered. This may be preferable to performing metaanalysis of several studies of different qualities, especially if the majority are of C grade, as is likely often to be the case. Different weights can also be chosen to reflect the quality of the articles included in the metaanalysis. However, in our study this would have only limited effects on the global estimate, as the CI of the EuBIVAS estimate is narrow compared with those of the 2 C studies; this has the greatest impact when the number of included estimates is small. For example, if keeping the weight for the C articles unchanged, reducing the weight for the A grade from 4 to 3 gives a CV_I estimate of 10.5%, whereas an increase to 10 gives a CV_I estimate of 9.5%, i.e., close to the original EuBIVAS estimate. An alternative weighting approach could be to use the weights for each QI on the checklist to calculate a mean BIVAC quality grade that could be used as weight in the metaanalysis. Results from the metaanalysis of BIVAC-appraised publications for the other measurands included in this review will be made available in a new BV database on the EFLM website.

Conclusions

The BIVAC enables review of studies on BV to critically appraise and classify them with regard to study design, preanalytical handling, analytical methods, and statistical analysis. Our data from a rigorous study of the application of BIVAC to a large volume of published studies indicate that many BV studies omit, or fail to address, essential detail regarding the BIVAC QIs. This may affect the reliability of the associated estimates. The BIVAC has resulted from an exhaustive discussion and testing by a group of experts in the field. Further iterations may be required in the future for the BIVAC to reflect experiences from its application, input from the laboratory community, and changes in the type of BV studies that are published. Presently, the BIVAC not only enables a retrospective assessment of published studies but also can serve as a guide to those aspiring to deliver future studies. Thus, it provides the potential to drive up quality of future studies and as such may be considered as an initiative that may find an application for BV studies in much the same way as STARD (35) and STROBE (36) have found traction in the context of publications regarding diagnostic testing.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors' Disclosures or Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

Employment or Leadership: None declared.

Consultant or Advisory Role: None declared.

Stock Ownership: None declared.

Honoraria: None declared.

Research Funding: None declared.

Expert Testimony: None declared.

Patents: None declared.

Other Remuneration: Siemens Healthineers and Roche Diagnostics have provided funding for travel and accommodation for meetings in the EFLM WG-BV and TFG-BVD.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or final approval of manuscript.

Acknowledgments: The authors thank Callum Fraser, Per Hyltoft Petersen, Fernando Cava, and Mauro Panteghini for valuable comments in the development of the BIVAC. Additionally, they thank the Analytical Quality Commission of the Spanish Society of Laboratory Medicine for hosting the EFLM WG-BV and TFG-BVD meetings, and Siemens Healthineers and Roche Diagnostics for providing funding for travel and accommodation.

References

1. Fraser CG. Biological variation: from principles to practice. Washington (WA): AACC Press; 2001. p. 18–22.
2. Minchinella J, Ricos C, Perich C, Fernández-Calle P, Álvarez V, Doménech MV, et al. Biological variation database and quality specifications for imprecision, bias and total error (desirable and minimum). <http://www.westgard.com/biodatabase1.htm#1> (Accessed September 2017).
3. Ricos C, Alvarez V, Cava F, García-Lario JV, Hernandez A, Jimenez CV, et al. Current databases on biological variation: pros, cons and progress. *Scand J Clin Lab Invest* 1999;59:491–500.
4. Perich C, Minchinella J, Ricos C, Fernandez-Calle P, Alvarez V, Doménech MV, et al. Biological variation database: structure and criteria used for generation and update. *Clin Chem Lab Med* 2015;53:299–305.
5. Aarsand AK, Roraas T, Sandberg S. Biological variation—reliable data is essential. [Editorial]. *Clin Chem Lab Med* 2015;53:153–4.
6. Carobene A. Reliability of biological variation data available in an online database: need for improvement. *Clin Chem Lab Med* 2015;53:871–7.
7. Fraser CG, Sandberg S. Biological variation. In: Rifai N, Horvath AR, Wittwer CT, editors. *Tietz textbook of clinical chemistry and molecular biology*. 6th Ed. St. Louis (MO): Elsevier; 2017. p. 157–70.
8. Panteghini M, Sandberg S. Defining analytical performance specifications 15 years after the Stockholm conference. [Editorial]. *Clin Chem Lab Med* 2015;53:829–32.
9. European Federation of Clinical Chemistry and Laboratory Medicine Task and Finish Group Biological Variation Database. www.eflm.eu/site/page/a/1084 (Accessed September 2017).
10. European Federation of Clinical Chemistry and Laboratory Medicine Biological Variation Working Group. www.eflm.eu/site/page/a/1148 (Accessed September 2017).
11. Bartlett WA, Braga F, Carobene A, Coskun A, Prusa R, Fernandez-Calle P, et al. A checklist for critical appraisal of studies of biological variation. *Clin Chem Lab Med* 2015;53:879–85.
12. Simundic AM, Kackov S, Miler M, Fraser CG, Petersen PH. Terms and symbols used in studies on biological variation: the need for harmonization. *Clin Chem* 2015;61:438–9.
13. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to meta-analysis*. Chichester (UK): John Wiley & Sons Ltd.; 2009. p. 69–75.
14. Burdick RK, Graybill FA. Confidence intervals on variance components. *Statistics: textbooks and monographs*, Vol. 127. New York (NY): Marcel Dekker; 1992. p. 78–115.
15. Cormen TH, Leiserson CE, Rivest RL, Stein C. *Introduction to algorithms*. Cambridge (MA): The MIT Press, Massachusetts Institute of Technology; 1989. 194 p.
16. Ropen source language and environment for statistics. www.cran.rstudio.com/web/packages/matrixStats/index.html (Accessed September 2017).
17. Tu D, Shao J. *The jackknife and bootstrap*. 1st Ed. New York (NY): Springer Series in Statistics; 1995.
18. Bailey D, Bevilacqua V, Colantonio DA, Pasic MD, Perumal N, Chan MK, Adeli K. Pediatric within-day biological variation and quality specifications for 38 biochemical markers in the CALIPER cohort. *Clin Chem* 2014;60:518–29.
19. Statland BE, Winkel P, Bokelund H. Factors contributing to intra-individual variation of serum constituents. 1. Within-day variation of serum constituents in healthy subjects. *Clin Chem* 1973;19:1374–9.
20. Qi Z, Chen Y, Zhang L, Ma X, Wang F, Cheng Q, et al. Biological variations of thirteen plasma biochemical indicators. *Clin Chim Acta* 2016;452:87–91.
21. Winkel P, Statland BE, Bokelund H. Factors contributing to intra-individual variation of serum constituents: 5. Short-term day-to-day and within-hour variation of serum constituents in healthy subjects. *Clin Chem* 1974;20:1520–7.
22. Hölzel WG. Intra-individual variation of analytes in serum from patients with chronic liver diseases. *Clin Chem* 1987;33:1133–6.
23. Hölzel WG. Intra-individual variation of some analytes in serum of patients with insulin-dependent diabetes mellitus. *Clin Chem* 1987;33:57–61.
24. Pineda-Tenor D, Laserna-Mendieta EJ, Timon-Zapata J, Rodelgo-Jimenez L, Ramos-Corral R, Recio-Montealegre A, Reus MG. Biological variation and reference change values of common clinical chemistry and haematologic laboratory analytes in the elderly population. *Clin Chem Lab Med* 2013;51:851–62.
25. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group. <http://www.gradeworkinggroup.org/> (Accessed September 2017).
26. Lenters-Westra E, Roraas T, Schindhelm RK, Slingerland RJ, Sandberg S. Biological variation of hemoglobin a1c: consequences for diagnosing diabetes mellitus. *Clin Chem* 2014;60:1570–2.
27. Weykamp CW, Mosca A, Gillery P, Panteghini M. The analytical goals for hemoglobin a(1c) measurement in IFCC units and National Glycohemoglobin Standardization program units are different. *Clin Chem* 2011;57:1204–6.
28. Carobene A, Braga F, Roraas T, Sandberg S, Bartlett WA. A systematic review of data on biological variation for alanine aminotransferase, aspartate aminotransferase and gamma-glutamyl transferase. *Clin Chem Lab Med* 2013;51:1997–2007.
29. Roraas T, Petersen PH, Sandberg S. Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals. *Clin Chem* 2012;58:1306–13.
30. Biosca C, Ricos C, Jimenez CV, Lauzurica R, Galimany R. Model for establishing biological variation in non-healthy situations: renal posttransplantation data. *Clin Chem* 1997;43:2206–8.
31. Roraas T, Stove B, Petersen PH, Sandberg S. Biological variation: the effect of different distributions on estimated within-person variation and reference change values. *Clin Chem* 2016;62:725–36.
32. Braga F, Ferraro S, Ieva F, Paganoni A, Panteghini M. A new robust statistical model for interpretation of differences in serial test results from an individual. *Clin Chem Lab Med* 2015;53:815–22.
33. Burdick RK, Borror C, Montgomery D. *Design and analysis of gauge R&R studies*. 1st Ed. Philadelphia (PA): Society for Industrial Applied Mathematics (SIAM); 2005. p. 58–63.
34. Carobene A, Roraas T, Solvik UO, Sylte MS, Sandberg S, Guerra E, et al. Biological variation estimates obtained from 91 healthy study participants for 9 enzymes in serum. *Clin Chem* 2017;63:1141–50.
35. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Clin Chem* 2015;61:1446–52.
36. Strengthening the reporting of observational studies in epidemiology (STROBE). www.strobe-statement.org/index.php?id=strobe-home (Accessed September 2017).