Method agreement analysis and interobserver reliability of the ISTH proposed definitions for effective hemostasis in management of major bleeding

Abdoellakhan, Rahat A.; Beyer-Westendorf, Jan; Schulman, Sam; Sarode, Ravi; Meijer, Karina; Khorsand, Nakisa

Published in:
JOURNAL OF THROMBOSIS AND HAEMOSTASIS

DOI:
10.1111/jth.14388

Journal of Thrombosis and Haemostasis, **17**: 499–506 DOI: 10.1111/jth.14388

ORIGINAL ARTICLE

Method agreement analysis and interobserver reliability of the ISTH proposed definitions for effective hemostasis in management of major bleeding

RAHAT A. ABDOELLAKHAN,* JAN BEYER-WESTENDORF,†‡ SAM SCHULMAN,§ RAVI SARODE,¶ KARINA MEIJER* and NAKISA KHORSAND**

**Department of Hematology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; †Thrombosis Research Unit, Department of Medicine I, Division Hematology, University Hospital "Carl Gustav Carus" Dresden, Dresden, Germany; ‡Kings Thrombosis Service, Department of Hematology, Kings College London, London, UK; §Department of Medicine, McMaster University, Hamilton, Ontario, Canada; ¶Division of Transfusion Medicine and Hemostasis, Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, USA; and **Department of Pharmacy, OLVG, Amsterdam, the Netherlands*

To cite this article: Abdoellakhan RA, Beyer-Westendorf J, Schulman S, Sarode R, Meijer K, Khorsand N. Method agreement analysis and interobserver reliability of the ISTH proposed definitions for effective hemostasis in management of major bleeding. *J Thromb Haemost* 2019; **17**: 499–506.

Essentials

- In 2016 the SSC proposed definitions for effective hemostasis in management of major bleeding.
- To validate these definitions, we studied the use in three large anticoagulant-reversal studies.
- Method agreement analysis and interobserver reliability showed at least acceptable agreement.
- Recommendations were made, advising use of the definition in hemostatic effectiveness studies.

Summary. *Introduction:* In 2016 the Scientific and Standardization Subcommittee (SSC) on Control of Anticoagulation of the International Society on Thrombosis and Haemostasis (ISTH) proposed criteria to evaluate the effectiveness of anticoagulant reversal in major bleeding management. Testing and validation of these criteria are required. *Objective:* To investigate the method agreement, interobserver reliability and applicability of the ISTH proposed definitions for hemostatic effectiveness. *Methods:* Patient data from three anticoagulant-antidote studies were used for hemostatic effectiveness assessment using the ISTH-proposed definitions and clinical opinion. For every patient a case document was produced. For each cohort, four adjudicators were asked to assess the hemostatic effectiveness independently on a case-by-case basis. Agreement between the two methods of hemostatic effectiveness assessment was calculated using Cohen's kappa (κ), with a calculated sample size of at least 73 cases. *Results:* The full dataset consisted of 116 cases, resulting in 464 assessments. Method agreement in outcome was observed in 364 of 464 assessments (78.5%), resulting in κ of 0.634 (95% CI: 0.575–0.694), or "substantial agreement." Interobserver reliability analysis of the proposed definitions computed an overall agreement of 54.2% with κ of 0.312 ("fair agreement"). *Discussion:* Method agreement analysis shows that the conclusions drawn using the ISTH definitions have "substantial agreement" with clinical opinion. Interobserver reliability analysis demonstrated acceptable agreement. In-depth analysis provided minor opportunities for further improvement and correct application of the definition. The definition is recommended to be used in all future studies evaluating hemostatic effectiveness, taking the suggested recommendations into account.

Keywords: anticoagulants; bleeding; hemostasis; outcome assessment; prothrombin complex concentrates.

Correspondence: Rahat A. Abdoellakhan, Department of Hematology, University Medical Center Groningen, Hanzeplein 1, 9713 GZ Groningen, the Netherlands
Tel.: +31 50 361 0225
E-mail: r.a.abdoellakhan@umcg.nl

Received: 6 September 2018
Manuscript handled by: M. Carrier
Final decision: F. R. Rosendaal, 14 January 2019

Introduction

A common challenge for studies investigating the hemostatic effectiveness of an intervention in bleeding patients on anticoagulants consists of defining and measuring clinical outcome. In the absence of a standardized definition, studies evaluating the effect of antidotes for oral anticoagulants often seek to define surrogate laboratory

parameters such as improvement/normalization of international normalized ratio (INR), diluted thrombin time, or anti-factor Xa activity as primary measurement of anticoagulation reversal. Whenever clinical outcome is evaluated, it is usually with *ad hoc* formulated definitions assessing hemostatic effectiveness.

Recent landmark studies on reversal of anticoagulants highlight this problem. In 2013 Sarode *et al*. reported in a study evaluating prothrombin complex concentrate (PCC) a definition for hemostatic effectiveness formulated in consultation with the US Food and Drug Administration (FDA) [1]. With input from the regulatory authority they designed a rational definition for hemostatic effectiveness. The study evaluating idarucizumab by Pollack *et al*. in 2017 reported clinical outcome by assessment of the extent of bleeding and hemodynamic stability at multiple time points [2]. The third study of note is the andexanet alfa study by Connolly *et al*. in 2016, which reported the use of an adapted version of the Sarode criteria [3].

The lack of standardized definitions introduces bias and hampers comparison between treatments and studies. This was first acknowledged in a systematic review comparing PCC dosing strategies in 2015 [4], which prompted the Scientific and Standardization Subcommittee (SSC) on Control of Anticoagulation of the International Society on Thrombosis and Haemostasis (ISTH) to approach the problem. As a result, definitions were proposed for assessment of effectiveness of major bleeding management in 2016 [5], prepared by a working group consisting of the same authors as the current project.

The recently proposed definitions formulate, for each specific bleeding type, criteria that should be met in regard to the hemostatic treatment outcome as "effective." A schematic summary of the proposed criteria per bleeding type is given in Fig. 1; full details can be found in [5].

However, the preceding definitions only represent an expert consensus so far, and testing and validation of these criteria are therefore required. Agreement should be determined between the new method and the current clinical gold standard, which is the opinion of the physician involved in the bleeding management at the bedside. Furthermore, interobserver reliability should be determined and limitations in applicability of the new method need to be identified and resolved.

The current study seeks to test and validate the proposed definitions to increase understanding of the feasibility and limitations of this assessment tool and ultimately provide a justification for use in future clinical trials, but also in clinical practice.

Methods

Aims

The primary aim for this study was to investigate the method agreement of hemostatic effectiveness assessment using the ISTH-proposed definitions and clinical opinion. Furthermore, interobserver variability of the ISTH-proposed definitions was analyzed and applicability of the proposed criteria was studied.

Cases and adjudication

For hemostatic effectiveness assessments, we used patient data from three anticoagulant-reversal studies or registries that the authors had access to. The studies were regarded as three separate cohorts throughout the current project, each with its own specific characteristics. A summary of important details on the three cohorts is given in Table 1. All eligible patients who

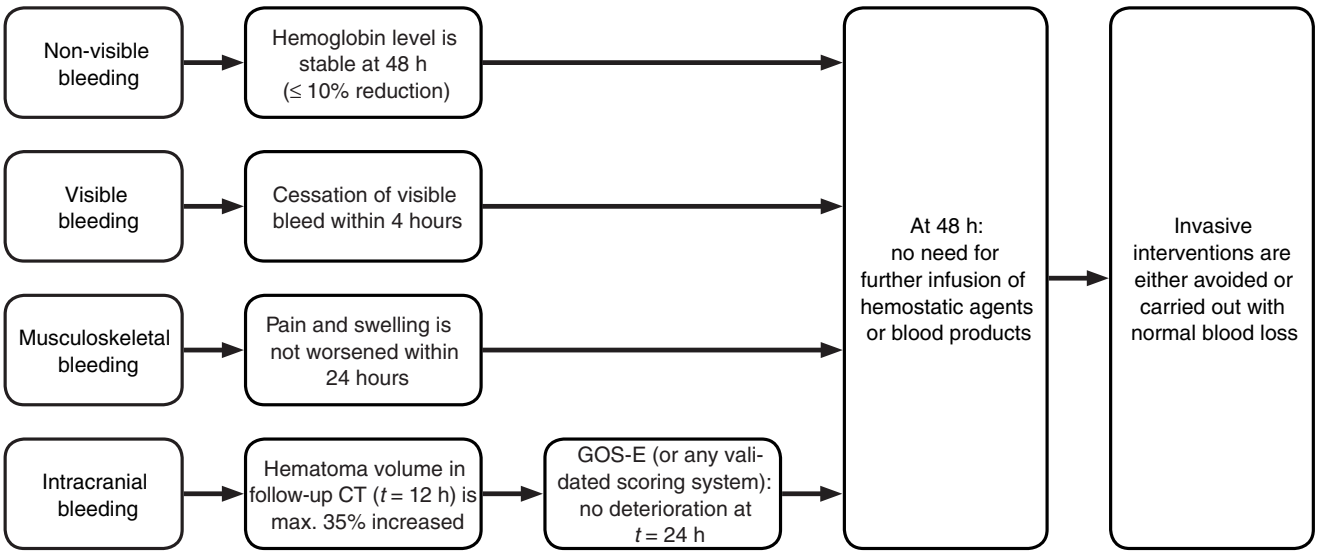

Fig. 1. Simplified schematic representation of ISTH-proposed definitions for effective hemostasis in management of major bleeding. GOS-E, Extended Glasgow Outcome Scale.

Table 1 Cohort characteristics of cases included

	Cohort A [6]	Cohort B [7–10]	Cohort C [11]
Cases (*n*)	40	19	57
Type of study	Multicenter RCT	Multicenter registry	Multicenter cohort
Hemostatic effectiveness predefined?	Yes, ISTH definition	No	Yes, definition from [1]
Anticoagulant	VKA	NOACs	NOACs
Studied reversal agent	4F-PCC	4F-PCC	4F-PCC
Country	The Netherlands	Germany	Canada

4F-PCC, 4-factor prothrombin complex concentrate; NOAC, non-vitamin K antagonist oral anticoagulant; RCT, randomized controlled trial; VKA, vitamin K antagonist.

were available at the time of this study from the underlying ongoing projects of cohorts A and B were included, whereas for cohort C data collection stopped after including the first 57 consecutive patients, meeting sufficient sample size.

For every patient within these cohorts a case document was produced. For cohorts A and B, this included admission and discharge notes, all progress notes, laboratory data, transfusion data, medication log, and imaging reports. For cohort C case summaries were composed, describing admission, progress, imaging and discharge notes, and relevant laboratory and transfusion data. Adjudicators were blinded to details of the hemostatic agent of interest when the original trial intervened in dose or regimen of that hemostatic agent.

For each cohort, four adjudicators were asked to assess the hemostatic effectiveness on a case-by-case basis independently. Two of four adjudicators were part of the SSC working group that had formulated the ISTH criteria, representing the "working group" observers. The other two were physicians experienced in the assessment of bleeding, but not previously involved with the ISTH criteria, representing "naïve" observers.

A case assessment form was developed in which the adjudicator was first asked to assess the hemostatic effectiveness according to the adjudicator's clinical opinion, the current gold standard method in clinical practice. Subsequently the adjudicator was asked to reassess the hemostatic effectiveness using the new ISTH-proposed criteria.

Of note, a necessary adjustment, which was erroneously missing in the published proposed definitions [5], was made to the definition beforehand: it was allowed to replace the Extended Glasgow Outcome Scale (GOS-E) with other validated scoring systems to assess neurologic outcome in intracranial hemorrhage (ICH). For this specific project, we chose to allow the Glasgow Coma Scale (GCS) to be used as an alternative when GOS-E was missing. In case of intraobserver discrepancy between clinical opinion and assessment using the ISTH definition, or in case of non-assessability, a possible explanation was requested.

Analysis and statistics

Sample size calculation was based on the Cohen's κ test used for method agreement analysis in the full dataset, determining the chance-adjusted agreement between the two methods of hemostatic effectiveness assessment [12]. A previously described approach was used to calculate the sample size [13]. Assumptions for calculation were made, based on three possible outcomes (i.e., effective, non-effective, and not assessable) and four adjudicators for every case [14]. Anticipated κ was set at 0.8, maximum confidence interval width at 0.2, an estimated proportion of categories at 0.2, 0.3, and 0.5, with $\alpha = 0.05$ and $\beta = 0.20$. This resulted in a minimum sample size of 73 cases to be assessed by four observers, totaling 292 assessments.

Interobserver reliability of the four observers was calculated using the free marginal Fleiss' κ statistic to determine multiple (>2) rater chance-adjusted agreement [15–17]. A subgroup analysis for interobserver agreement was performed to exclude bias introduced by non-assessable cases, by excluding cases that one or more of the observers had rated as not assessable according to the ISTH definitions.

The applicability of the definition consisted of analysis of (i) interobserver agreement in bleeding type, (ii) outcome assessment analysis per bleeding type and (iii) in-depth analysis of non-assessability of cases. For analysis of interobserver agreement in bleeding type, Fleiss' κ statistic was used. Cases with discrepancy in bleeding type between adjudicators were analyzed for consequences in hemostatic effectiveness outcome.

For in-depth analysis of the non-assessability of cases, correct bleeding types were retrospectively (after completion of all assessments) assigned to every case by two members of the working group, with the help of a third if no consensus could be reached. Then cases were categorized per correct bleeding type and analyzed for proportion of assessments with corresponding bleeding type assignment by the adjudicators. Finally, the assessments with corresponding bleeding type assignment that concluded hemostatic effectiveness to be non-assessable were analyzed for frequency and reason for non-assessability.

Each analysis was performed on the full dataset at first and then, where applicable, for each cohort separately. Kappa values were interpreted using the definition of Landis and Koch; see Table 2 [18].

Results

The full dataset consisted of 116 cases, resulting in 464 assessments. In cohort A, four observers adjudicated 40 consecutive cases, totaling 160 assessments; cohort B

Table 2 Interpretation of kappa [18]

Kappa statistic	Agreement
< 0.00	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

totaled 76 assessments for 19 cases, and cohort C 228 assessments for 57 cases.

Method agreement analysis

Agreement in outcome between methods in the full dataset was observed in 364 of 464 cases (78.5%), resulting in a Cohen's κ of 0.634 (95% CI: 0.575–0.694), or "substantial agreement." In detail, κ in cohort A was 0.669 (95% CI: 0.553–0.785), in cohort B 0.467 (95% CI: 0.322–0.611), and in cohort C 0.657 (95% CI: 0.577–0.737). A sensitivity analysis was performed limited to cohorts with hemostatic effectiveness in some way predefined, i.e., cohorts A and C, totaling 97 cases. This produced a Cohen's κ of 0.670 (95% CI: 0.553–0.785), or "substantial agreement," confirming the agreement found in the full dataset.

The sensitivity and specificity of assessment using the ISTH definition in the full dataset were 74.3% and 86.9%. Details of sensitivity and specificity per cohort are displayed in Table 3.

Interobserver reliability

Interobserver reliability analysis of the proposed definitions on the full dataset computed an overall agreement of 54.2% with a Fleiss free marginal chance-corrected agreement κ of 0.312 ("fair agreement"). For comparison purposes, interobserver reliability for assessment based on adjudicator's clinical opinion had an agreement of 69.0% with κ 0.534 ("good agreement"). Working group observers produced a slightly better κ when using the proposed definitions compared to naïve observers upon stratification (0.392 and 0.276, respectively). Finally, when focusing only on cohorts with hemostatic effectiveness predefined (cohorts A and C), agreement was demonstrated to be slightly better than in the full dataset, with agreement being 57.2% and κ being 0.358 ("fair agreement"). Kappa values stratified per cohort and observers are displayed in Table 4.

For the subgroup analysis, 7 cases in cohort A (18%) were rated not assessable by at least one of the observers when using the ISTH-proposed definitions. In the same manner 16 cases (84%) were not assessable in cohort B and 32 cases (56%) in cohort C. Consequently, κ was 0.333 in cohort A and 0.620 in cohort C, while for cohort B κ was not calculated because of a low number of remaining cases. Thirty-seven of the total of 55 cases (67%) that were not assessable when using the ISTH-proposed definitions were also not assessable by the adjudicator's clinical opinion.

Applicability of the definition

An in-depth analysis per bleeding type was performed, in which further analysis of the use and non-assessability of the ISTH definition was performed. For this analysis only, two members of the working group assigned correct bleeding types to the cases in retrospect. Consensus was reached for all cases.

Figure 2 displays the distribution of assessed outcomes specified per bleeding type. Here the absence of false positives is demonstrated, except in two assessments in the non-visible bleed type category (1%). Both were caused by evident adjudicator error, i.e., severe ongoing blood loss in one case and a recurrent bleed in the other, both of which should have resulted in non-effective hemostasis if ISTH-proposed definitions were followed correctly. False negative rates can also be read from Fig. 2 to be less than 9% in each bleeding type category except in musculoskeletal bleeds, in which the rate was 17%. Most common reasons for false negatives were found to be cessation of bleeding according to clinical opinion, but not

Table 3 Intraobserver, intermethod agreement expressed in percentage agreement and Cohen's kappa of hemostatic effectiveness determination using the ISTH definitions and clinical opinion. Sensitivity and specificity of the ISTH definition are also displayed

	Full dataset	Cohort A	Cohort B	Cohort C
Agreement (%)	364/464 (78.5%)	136/160 (85%)	49/76 (64.5%)	179/228 (78.5%)
Kappa (95% CI)	0.634 (0.575–0.694)	0.669 (0.553–0.785)	0.467 (0.322–0.611)	0.657 (0.577–0.737)
Sensitivity	74.3	83.3	55.3	72.9
Specificity	86.9	94.1	68.8	87.5

CI, confidence interval.

Table 4 Interobserver reliability, expressed as % overall agreement and chance corrected agreement (Fleiss' kappa)

	Full dataset	Cohort A	Cohort B	Cohort C
Agreement (%)	54.2%	60.8%	38.6%	54.7%
Kappa	0.312	0.413	0.079	0.320
Kappa (naive observers only)	0.276	0.363	0.211	0.237
Kappa (working group observers only)	0.392	0.438	0.368	0.368

A Non visible bleeding

ISTH \ Clinical opinion	Effective	Non-effective	Not assessable
Effective	98	2	2
Non-effective	17	48	4
Not assessable	9	3	10

B Visible bleeding

ISTH \ Clinical opinion	Effective	Non-effective	Not assessable
Effective	42	0	0
Non-effective	6	22	1
Not assessable	3	0	4

C Musculoskeletal bleeding

ISTH \ Clinical opinion	Effective	Non-effective	Not assessable
Effective	19	0	0
Non-effective	5	2	0
Not assessable	1	1	2

D Intracranial bleeding

ISTH \ Clinical opinion	Effective	Non-effective	Not assessable
Effective	61	0	0
Non-effective	5	34	1
Not assessable	30	10	19

Fig. 2. Contingency tables of hemostatic effectiveness assessment by clinical opinion and by ISTH-proposed definitions (ISTH), specified per bleeding type. (A) Non visible bleeding, (B) visible bleeding, (C) musculoskeletal bleeding, (D) intracranial bleeding.

meeting the ISTH-proposed criteria, in 50% and adjudicator error in 25%.

Table 5 gives an overview of the cases and their assessed bleeding types by the adjudicators, specified per correct bleeding type as assigned by the working group in retrospect. Furthermore, frequencies and reasons for non-assessability when using the ISTH definition are given for cases in which the assessed bleeding type was identical to the correct bleeding type. Most non-assessable cases originated from cohorts B and C. Most discrepancy between bleeding type as assessed by the adjudicator and the correct bleeding type was seen in the non-visible and visible bleeding types.

Interobserver agreement in bleeding type

Kappa values for interobserver agreement in bleeding type for cohorts A, B, and C were 0.606, 0.889, and 0.960, respectively. Full agreement between all observers was reached in 19/40 cases in cohort A, in 15/19 cases of cohort B, and in 53/57 cases in cohort C. Nearly all cases (23 of 25) that were assigned more than one bleeding type by observers were due to discrepancy in the discrimination between visible and non-visible bleeding types. These were 17 GI bleeds, 1 epistaxis combined with GI bleeding, 1 hematuria, 1 renal bleeding, 1 intraabdominal bleed, 1 vaginal bleed, and 1 case of hemoptysis.

In 10 of 23 cases that were assigned non-visible and visible bleed types, the bleed type assignment had no consequences for the hemostatic effectiveness conclusion. In 4 cases, however, there were consequences for the conclusion, meaning that bleeding type specific questions (i.e., hemoglobin drop at 48 h for non-visible bleeds or cessation at 4 h for visible bleeds) resulted in contrasting answers, leading to different conclusions in outcome. For the remaining cases it was inconclusive whether bleeding type assignment had consequences for the outcome.

Discussion

This study reports on the applicability and reliability of the ISTH-proposed definitions for assessment of the

Table 5 Bleeding type assessment distribution of cases categorized per correct bleeding type and frequencies with reasons of non-assessable cases when using the ISTH-proposed definition for hemostatic effectiveness assessment

	Cases (*n*)	Assessments (*n*)	Assessed bleeding type (*n*)	Correctly assessed bleeding type and ISTH not assessable (*n*)
ICH	41	164	ICH: 160 Non-visible: 3 Visible: 1	59 (37%) • Follow-up CT not assessable: 55 • GOS-E & GCS not assessable: 60
Musculoskeletal	8	32	Musculoskeletal: 30 Non-visible: 1 Not specified: 1	4 (13%) • Pain & swelling not assessable: 7 • In cohort B & C: 6
Non-visible	58	232	Non-visible: 189 Visible: 41 Not specified: 2	23 (12%)* • Hemoglobin not assessable: 22 • In cohort B & C: 19
Visible	9	36	Visible: 32 Non-visible: 4	6 (8%)* • Cessation not assessable: 6 • In cohort B & C: 6

GCS, Glasgow Coma Scale; GOS-E, Extended Glasgow Outcome Scale; ICH, intracranial hemorrhage. *For non-visible and visible bleeds, assessments were pooled because of large interobserver variability in bleeding type.

hemostatic effectiveness of anticoagulant reversal. The results of the method agreement analysis show that the conclusions drawn using the ISTH definitions have, to our expectations, "substantial agreement" with clinical opinion. This implies that the systematic, predefined approach of the proposed definitions has face value; it produces a similar and thereby acceptable outcome to clinical opinion.

In-depth analysis of the cases demonstrated the near-absence of false positives for all bleeding types. False negatives were for all bleed types below 9% except for musculoskeletal bleeds, in which a rate of 17% was found. The main reasons for false negatives were adjudicator error in 25% and, more importantly, clinically evident cessation of bleeding without meeting the bleeding-type specific criteria in 50%.

The presence of false negatives was, however, expected considering the conservative nature of the definition. In light of this, compared to the total number of assessments, a false negative rate of 7%, with approximately a quarter due to adjudicator error, can be regarded as acceptable as broadening of the criteria will likely lead to a higher number of false positives. For musculoskeletal bleeds, however, broadening the criterion to "no worsening of pain and swelling" instead of "pain and swelling" would improve face value. The assessment of pain and swelling should furthermore be predefined in the protocol of prospective studies. The false negative rates caused by adjudicator error could be resolved by instruction of the adjudicator and the use of more than one adjudicator.

Further incentive for adjustment of criteria arose from analysis of non-assessable cases. The cases rated most frequently as not assessable were from cohorts B and C, in which the definition was applied after data collection, i.e., *post hoc*. For cohort A, which was predefined to collect data relevant to the definition, the parameters required by the definition seemed feasible. In this predefined cohort ICH was, however, excluded so the question remains how feasible repeat CT and/or GOS-E or GCS scoring is for this type of cohort. In the *post hoc* setting it appears that the specifically required parameters at fixed time points, on which the definition depends, are often lacking. Consequently, in prospective data collection, the time points for repeat CT and/or GOS-E or GCS should be predefined in study protocols.

Another difference in application of the definitions between predefined and *post hoc* cohorts was identified in the interobserver reliability analysis. Although "fair agreement" for the proposed definition in the full dataset is acceptable, it suggests that the systematic and predefined approach still leaves some room for interpretation and disagreement between adjudicators. Stratification between cohorts clarified that this is especially the case if the data required by the definition is missing or not obvious enough. As a result interobserver reliability in cohort A, in which assessment criteria were predefined, showed 'good agreement', while *post hoc* cohorts B and C demonstrated only "slight agreement" and "fair agreement."

Last, interobserver agreement in bleeding type revealed that there is considerable variation in defining a bleeding to be visible or non-visible. This was also concluded from the in-depth analysis per bleeding type in Table 4. While it was obvious to classify intracranial bleeds and musculoskeletal bleeds as such correctly, it appeared to be less obvious to identify the rest of the bleeds to either visible or non-visible, which was especially common in bleedings that usually receive endoscopic diagnosis and/or treatment: gastrointestinal bleeding, hemoptysis, and hematuria. For standardization purposes, the current ISTH definition could provide more clarification on how to categorize such bleeding events as either visible or non-visible bleeds.

In the spirit of the development of definitions, visible bleeds were meant to be classified as such when the focus of the bleeding is directly visible (e.g., skin surface, visible mucosal bleed [oral/nose/anal]) or is located in a compartment in which blood cannot be occult for longer periods (e.g., hemoptysis, hematuria). Bleeds with non-visible focus that cannot be classified as musculoskeletal or intracranial bleeds (e.g., occult hemoglobin/blood loss) or bleeds located in compartments that could store blood for longer periods, should be classified as non-visible bleeds (e.g., GI bleeds, intraabdominal bleeds, parenchymal bleed). For these bleeding events, the course of hemoglobin levels is the most appropriate clinical way to assess hemostatic effectiveness during follow-up. Exact predefined guidance for data collection is recommended for future prospective studies.

The current work was performed with a large number of observers with diverse expertise and experience with these definitions. This benefited the applicability of study results with respect to adjudicators. The use of multiple cohorts stemming from different studies with unique characteristics, focusing on hemostatic interventions for various anticoagulants, contributed even further to the applicability of results in general. A disadvantage of the data was the lack of ICH cases in a predefined setting (cohort A).

Based on these validation results, the current ISTH definitions can be recommended to be used as standard for assessment of hemostatic effectiveness. On the basis of in-depth analysis, we recommend taking the following into account when using the definition:

- For prospective studies, design the study to promote the collection of parameters at the specified time points as required by the ISTH definitions.
- Use two or more adjudicators and have cases adjudicated independently with consensus forming after discussion.
- Make sure that adjudicators read and understand definitions and assessment criteria.

- For the intracranial bleeding type, it is advised (as was erroneously missing in the first publication) to allow replacement of GOS-E with any validated scoring system to assess neurologic outcome if GOS-E is not routinely collected (especially in *post hoc* settings). We would recommend GCS as a valid alternative.
- Prespecify precisely the categorization of non-visible and visible bleeds.

Conclusions

In conclusion, the ISTH-proposed definitions for effective hemostasis in management of major bleeding were validated for use in datasets containing the parameters needed to evaluate the criteria of the definition. The definition demonstrated good method agreement and fair interobserver agreement. In-depth analysis provided recommendations to improve application of these definitions further. These definitions are recommended to be used as standard for assessing hemostatic effectiveness in all future studies evaluating management of major bleeding, taking the formulated recommendations into account.

Addendum

R. A. Abdoellakhan, J. Beyer-Westendorf, S. Schulman, R. Sarode, K. Meijer, and N. Khorsand were involved in the study design. Data were collected by J. Beyer-Westendorf, S. Schulman, K. Meijer, and N. Khorsand. R.A. Abdoellakhan, K. Meijer, and N. Khorsand analyzed the data and wrote the concept version of the manuscript. All authors reviewed, contributed to, and approved the final version of the manuscript.

Acknowledgments

We thank our colleagues who contributed significantly in their role of "naïve" observers: Sebastian Endig, MD, Heike Endig, MD, Margriet Piersma-Wichers, MD, Mark Harms, MD, Jasper van Miert, MD, and Hilde Hop, MD.

Disclosure of Conflict of Interests

K. Meijer reports travel support, speaker fees, or consulting fees from Baxter, Bayer, Sanquin, Pfizer, Boehringer Ingelheim, BMS, Aspen, and Uniqure outside the submitted work. R. Sarode reports personal fees from CSL Behring, Octapharma, and Portola outside the submitted work. J. Beyer-Westendorf reports personal fees from Portola during the conduct of the study; grants and personal fees from Bayer and Daiichi Sankyo; grants from Boehringer Ingelheim, Pfizer; and personal fees from Janssen, outside the submitted work.

References

1 Sarode R, Milling TJ Jr, Refaai MA, Mangione A, Schneider A, Durn BL, Goldstein JN. Efficacy and safety of a 4-factor prothrombin complex concentrate in patients on vitamin K antagonists presenting with major bleeding: a randomized, plasma-controlled, phase IIIb study. *Circulation* 2013; **128**: 1234–43.
2 Pollack CV Jr, Reilly PA, van Ryn J, Eikelboom JW, Glund S, Bernstein RA, Dubiel R, Huisman MV, Hylek EM, Kam CW, Kamphuisen PW, Kreuzer J, Levy JH, Royle G, Sellke FW, Stangier J, Steiner T, Verhamme P, Wang B, Young L, *et al*. Idarucizumab for Dabigatran Reversal - Full Cohort Analysis. *N Engl J Med* 2017; **377**: 431–41.
3 Connolly SJ, Milling TJ Jr, Eikelboom JW, Gibson CM, Curnutte JT, Gold A, Bronson MD, Lu G, Conley PB, Verhamme P, Schmidt J, Middeldorp S, Cohen AT, Beyer-Westendorf J, Albaladejo P, Lopez-Sendon J, Goodman S, Leeds J, Wiens BL, Siegal DM, *et al*. ANNEXA-4 Investigators. Andexanet alfa for acute major bleeding associated with factor Xa inhibitors. *N Engl J Med* 2016; **375**: 1131–41.
4 Khorsand N, Kooistra HA, van Hest RM, Veeger NJ, Meijer K. A systematic review of prothrombin complex concentrate dosing strategies to reverse vitamin K antagonist therapy. *Thromb Res* 2015; **135**: 9–19.
5 Khorsand N, Majeed A, Sarode R, Beyer-Westendorf J, Schulman S, Meijer K; Subcommittee on Control of Anticoagulation. Assessment of effectiveness of major bleeding management: proposed definitions for effective hemostasis: communication from the SSC of the ISTH. *J Thromb Haemost* 2016; **14**: 211–4.
6 Abdoellakhan RA, Khorsand N, Van Hest RM, Veeger N, Ter Avest E, Ypma PF, Faber LM, Meijer K. Randomised controlled trial protocol to evaluate a fixed dose prothrombin complex concentrate against the variable dose in vitamin K antagonist related bleeding (PROPER3). *BMJ Open* 2018; **8**: e020764.
7 Beyer-Westendorf J, Förster K, Pannach S, Ebertz F, Gelbricht V, Thieme C, Michalski F, Köhler C, Werth S, Sahin K, Tittl L, Hänsel U, Weiss N. Rates, management, and outcome of rivaroxaban bleeding in daily care: results from the Dresden NOAC registry. *Blood* 2014; **124**: 955–62.
8 Beyer-Westendorf J, Ebertz F, Förster K, Gelbricht V, Michalski F, Köhler C, Werth S, Endig H, Pannach S, Tittl L, Sahin K, Daschkow K, Weiss N. Effectiveness and safety of dabigatran therapy in daily-care patients with atrial fibrillation. Results from the Dresden NOAC Registry. *Thromb Haemost* 2015; **113**: 1247–57.
9 Hecker J, Marten S, Keller L, Helmert S, Michalski F, Werth S, Sahin K, Tittl L, Beyer-Westendorf J. Effectiveness and safety of rivaroxaban therapy in daily-care patients with atrial fibrillation: results from the Dresden NOAC Registry. *Thromb Haemost* 2016; **115**: 939–49.
10 Helmert S, Marten S, Mizera H, Reitter A, Sahin K, Tittl L, Beyer-Westendorf J. Effectiveness and safety of apixaban therapy in daily-care patients with atrial fibrillation: results from the Dresden NOAC Registry. *J Thromb Thrombolysis* 2017; **44**: 169–78.
11 Schulman S, Gross PL, Ritchie B, Nahirniak S, Lin Y, Lieberman L, Carrier M, Crowther MA, Ghosh I, Lazo-Langner A, Zondag M; Study Investigators. Prothrombin complex concentrate for major bleeding on factor Xa inhibitors: a prospective cohort study. *Thromb Haemost* 2018; **118**: 842– 51.
12 Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measur* 1960; **20**: 37–46.
13 Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. *Theriogenology* 2010; **73**: 1167–79.
14 Rotondi MA, Donner A. A confidence interval approach to sample size estimation for interobserver agreement studies with

multiple raters and outcomes. *J Clin Epidemiol* 2012; **65**: 778–84.
15 Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971; **76**: 378–82.
16 Warrens MJ. Inequalities between multi-rater kappas: advances in data analysis and classification. *Adv Data Anal Classif* 2010; **4**: 271–86.
17 Randolph JJ. Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa. Paper presented at the Joensuu University Learning and Instruction Symposium 2005. Joensuu, Finland; 2005 Oct 14.
18 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–74.