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 Haematology, 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


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.

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...
Table 1 Cohort characteristics of cases included

<table>
<thead>
<tr>
<th>Cohort (n)</th>
<th>Type of study</th>
<th>Hemostatic effectiveness predefined?</th>
<th>Anticoagulant</th>
<th>Studied reversal agent</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort B [7-10]</td>
<td>Multicenter registry</td>
<td>No</td>
<td>NOACs</td>
<td>4F-PCC</td>
<td>Germany</td>
</tr>
</tbody>
</table>

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 $\kappa$ 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 $\kappa$ 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 $z = 0.05$ 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’ $\kappa$ 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’ $\kappa$ 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]

<table>
<thead>
<tr>
<th>Kappa statistic</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.00</td>
<td>Less than chance agreement</td>
</tr>
<tr>
<td>0.01–0.20</td>
<td>Slight agreement</td>
</tr>
<tr>
<td>0.21–0.40</td>
<td>Fair agreement</td>
</tr>
<tr>
<td>0.41–0.60</td>
<td>Moderate agreement</td>
</tr>
<tr>
<td>0.61–0.80</td>
<td>Substantial agreement</td>
</tr>
<tr>
<td>0.81–1.00</td>
<td>Almost perfect agreement</td>
</tr>
</tbody>
</table>

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 $\kappa$ of 0.634 (95% CI: 0.575–0.694), or “substantial agreement.” In detail, $\kappa$ 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 $\kappa$ 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 for 57 cases.

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

Table 3  Intraobserver, intermethod agreement expressed in percentage 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

<table>
<thead>
<tr>
<th></th>
<th>Full dataset</th>
<th>Cohort A</th>
<th>Cohort B</th>
<th>Cohort C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement (%)</td>
<td>364/464</td>
<td>136/160</td>
<td>49/76</td>
<td>179/228</td>
</tr>
<tr>
<td>(Kappa (95% CI)</td>
<td>(78.5%)</td>
<td>(85%)</td>
<td>(64.5%)</td>
<td>(78.5%)</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.634</td>
<td>0.609</td>
<td>0.467</td>
<td>0.657</td>
</tr>
<tr>
<td>(0.553–0.785)</td>
<td>(0.322–0.611)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>74.3</td>
<td>83.3</td>
<td>55.3</td>
<td>72.9</td>
</tr>
<tr>
<td>Specificity</td>
<td>86.9</td>
<td>94.1</td>
<td>68.8</td>
<td>87.5</td>
</tr>
</tbody>
</table>

CI, confidence interval.

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

<table>
<thead>
<tr>
<th></th>
<th>Full dataset</th>
<th>Cohort A</th>
<th>Cohort B</th>
<th>Cohort C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement (%)</td>
<td>54.2%</td>
<td>60.8%</td>
<td>38.6%</td>
<td>54.7%</td>
</tr>
<tr>
<td>(Kappa)</td>
<td>0.312</td>
<td>0.413</td>
<td>0.079</td>
<td>0.320</td>
</tr>
<tr>
<td>(95% CI)</td>
<td>(0.276–0.363)</td>
<td>(0.211–0.237)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.392</td>
<td>0.438</td>
<td>0.368</td>
<td>0.368</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.869</td>
<td>0.941</td>
<td>0.688</td>
<td>0.875</td>
</tr>
</tbody>
</table>

© 2019 The Authors. *Journal of Thrombosis and Haemostasis* published by Wiley Periodicals, Inc. on behalf of International Society on Thrombosis and Haemostasis.
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

<table>
<thead>
<tr>
<th>Cases (n)</th>
<th>Assessments (n)</th>
<th>Assessed bleeding type (n)</th>
<th>Correctly assessed bleeding type and ISTH not assessable (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICH</td>
<td>41</td>
<td>164</td>
<td>59 (37%)&lt;br&gt;- Follow-up CT not assessable: 55&lt;br&gt;- GOS-E &amp; GCS not assessable: 60</td>
</tr>
<tr>
<td>Musculoskeletal</td>
<td>8</td>
<td>32</td>
<td>4 (13%)&lt;br&gt;- Pain &amp; swelling not assessable: 7&lt;br&gt;- In cohort B &amp; C: 6</td>
</tr>
<tr>
<td>Non-visible</td>
<td>58</td>
<td>232</td>
<td>23 (12%)&lt;br&gt;- Hemoglobin not assessable: 22&lt;br&gt;- In cohort B &amp; C: 19</td>
</tr>
<tr>
<td>Visible</td>
<td>9</td>
<td>36</td>
<td>6 (8%)&lt;br&gt;- Cessation not assessable: 6&lt;br&gt;- In cohort B &amp; C: 6</td>
</tr>
</tbody>
</table>

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.

© 2019 The Authors. *Journal of Thrombosis and Haemostasis* published by Wiley Periodicals, Inc. on behalf of International Society on Thrombosis and Haemostasis.
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 the 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 Unique 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
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.
14 Rotondi MA, Donner A. A confidence interval approach to sample size estimation for interobserver agreement studies with
18 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–74.