Classification and management of shoulder and elbow trauma
Bruinsma, W.E.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 2

The Kappa Paradox

1,2Wendy Bruinsma MD, 3Job N. Doornberg, MD PhD, 2David Ring MD PhD, 1J. Carel Goslings MD PhD, 3Niels W.L. Schep, MD PhD MSc

On behalf of the Science of Variation Group (all observers to date)

1Trauma Unit, Department of Surgery
Academic Medical Center Amsterdam, Amsterdam, The Netherlands
2Hand and Upper Extremity Service, Department of Orthopaedic Surgery, Harvard Medical School, Massachusetts General Hospital, Boston, MA, USA
3Orthopaedic Sports & Traumatology, Department of Orthopaedic Surgery
Academic Medical Center Amsterdam, Amsterdam, The Netherlands

Submitted
ABSTRACT

Introduction

Observer reliability studies for fracture classification systems evaluate agreement using Cohen's kappa as well as absolute agreement as outcome measures. Cohen's kappa is a chance-corrected measure of agreement and usually ranges between 0 (no agreement) and 1 (perfect agreement). Absolute agreement is the percentage of times observers agree on the matter they have to rate; it does not depend on the answer (yes/no) as long as they agree. Some studies report a high absolute agreement, but a relatively low kappa value, which is contradictory. This phenomenon is called the Kappa Paradox. The objective of this review is to explain the statistical phenomenon of the Kappa Paradox to the Orthopaedic Readership.

Explaining the Kappa Paradox

The kappa statistic corrects for the difference between the observed agreement (o) and the agreement expected by chance (c). The kappa statistic is denoted as (o – c)/(1-c). The higher 'c' is, the lower the kappa coefficient will be. However, the agreement expected by change is influenced by the distribution of positive and negative cases displayed to the observers. When there is a high amount of only positive or only negative cases, the observed agreement based on chance will increase, in turn lowering the Kappa value without influencing the absolute agreement.

In a published interobserver study, 107 observers were asked to score whether or not the humeral head was split based on CT scans. A low kappa value of 0.26 and high absolute agreement of 0.95 was found. Analysis of the case distribution showed that 14 cases had a humeral head split and 1 did not. This imbalance in case and answer distribution rendered low kappa values.

Conclusions

Kappa values are influenced by case and answer distribution. We advise investigators to balance their case distributions to ensure that the amount of agreement by chance is as low as possible.
BACKGROUND

Observer reliability for classification systems of fractures has been widely studied in orthopaedic literature\(^{(1-3, 5-8, 10, 11, 13, 15, 17)}\). A perfect classification system should: 1) guide treatment; 2) predict clinical outcome; and 3) facilitate comparison of clinical studies from different institutions and investigators\(^{(14)}\). Since its introduction by Cohen in 1960, kappa values have been used to define agreement in intra- and inter observer studies\(^{(4, 12, 16)}\).

The kappa value (\(\kappa\)) is a chance-corrected measure of agreement comparing the observed measure of agreement with the level of agreement expected by chance alone. The kappa value can be classified according to Landis and Koch\(^{(9)}\) indicating the quantity of agreement: \(\kappa\) less than 0.0 is “poor,” 0.0 to 0.20 “slight,” 0.21 to 0.40 “fair,” 0.41 to 0.60 “moderate,” 0.61 to 0.80 “substantial,” and greater than 0.80 “almost perfect”.

Studies regularly report the absolute agreement alongside their kappa values\(^{(1)}\). The absolute agreement is the percentage of times observers agree on the matter they have to rate; it does not depend on the answer (yes/no) as long as they agree. If this absolute agreement is e.g. 95%, this means that 95 out of a 100 ratings, observers agreed with each other. Sometimes a study may report a high absolute agreement, but a relatively low kappa value. This phenomenon that the observed percentage of agreement is high and the kappa value is lower than one might expect based on the observed percentage of agreement is called the Kappa Paradox.

The objective of this study is to explain the statistical phenomenon of the Kappa Paradox to the Orthopaedic Readership.

EXPLAINING THE KAPPA PARADOX

In order to explain the Kappa Paradox we will discuss two interobserver studies; one hypothetical and one published study\(^{(1)}\).

The hypothetical study consists of two observers rating 100 proximal humeral fractures and answering one question: Is the Greater Tuberosity displaced; ‘yes’ or ‘no’?

In the published study, 107 observers were asked 6 questions about the radiographs of 15 different proximal humerus fractures. Four of the six questions could be answered with a yes or no, rendering binary results: 1) Is the humeral head split?; 2) Is the Greater Tuberosity displaced?; 3) Is the arterial supply compromised?; and 4) Is the glenohumeral joint dislocated? The other two questions concerned the Neer and AO-classification. We will focus on questions 1 and 4, as well as the results from the published study to further illustrate the Kappa Paradox.

Hypothetical Study

The results of the hypothetical study are presented in Table 1. The percentage of agreement is 85 + 4 (the number of times the observers agreed on a displaced greater tuberosity added to the number of times they agreed there was not) divided by the total number of cases = (85 + 4)/100 = 89%.
The Kappa Paradox

If the answers of the two observers would be totally random, in some cases there will be agreement based on chance. To calculate the agreement based on chance we turn to the marginal values (these are the values to the side of the table). Looking at the positive results, observer 1 rated 91 cases as ‘yes the tuberosity was displaced’. The marginal value would then be $91/100 = 0.91$. For observer 2 the positive marginal value would be $90/100 = 0.90$. If the two observers would be completely independent, the amount of agreement on a positive result that should be expected by chance would be $0.91 \times 0.90 = 0.82$. The amount of agreement on a negative result that could be expected by chance would then be $= (9/100) \times (10/100) = 0.01$. The total expected agreement based on chance would be the amount of agreement on a negative result added to the amount of agreement on a positive result $= 0.82 + 0.01 = 0.83$.

The observed agreement is denoted as ‘o’ and the agreement expected by chance as ‘c’. The kappa statistic corrects for the difference between the observed agreement (o) and the agreement expected by chance (c) as a proportion of the difference between a perfect agreement and the agreement expected by chance: $1 – c$. The kappa statistic is therefore $(o – c) / (1-c)$. As illustrated above, the higher ‘c’ is, the lower the kappa coefficient will be.

In this hypothetical study, the agreement according to the kappa value of our two observers rating the displacement of the greater tuberosity in a binary fashion the kappa would be: $(0.91 – 0.83)/(1-0.83) = 0.47$, which is classified as moderate agreement according to Landis and Koch. The kappa value of 0.47 combined with the total agreement of 0.89 seems contradictory; this is called the Kappa Paradox.

Data Published Study

In the published study, question one was: “Is the humeral head split?”. Agreement of this question was ‘fair’ with a kappa of 0.26. This kappa value is contra-intuitive, as one may think that diagnosing a humeral head is split should not render such a low agreement. This is supported by the percentage of absolute agreement of 0.95 in this study (Table 2).

<table>
<thead>
<tr>
<th>Question</th>
<th>$\kappa$</th>
<th>Classification</th>
<th>Percentage of overall agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the head split?</td>
<td>0.26</td>
<td>Fair</td>
<td>0.95</td>
</tr>
<tr>
<td>Is the glenohumeral joint dislocated?</td>
<td>0.12</td>
<td>Fair</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 2. Kappa values vs. percentage of overall agreement in the published study
We proceeded to review the case distribution. In this study, cases were distributed in a way that 14 cases had a head split, and one did not (Table 3). We altered the results so the cases would be symmetrically distributed; e.g. 50% of the cases did have a displaced greater tuberosity and 50% of the cases did not (Table 4).

<table>
<thead>
<tr>
<th>Case Category</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>38</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>44</td>
</tr>
<tr>
<td>3</td>
<td>39</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>38</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>38</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>41</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>41</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>41</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>41</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>110</td>
<td>580</td>
</tr>
</tbody>
</table>

Table 3. Distribution of answers to: “Is the head split?”

<table>
<thead>
<tr>
<th>Case Category</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>44</td>
</tr>
<tr>
<td>2</td>
<td>44</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>4</td>
<td>45</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>6</td>
<td>43</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>8</td>
<td>44</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>44</td>
</tr>
<tr>
<td>10</td>
<td>44</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>12</td>
<td>44</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>14</td>
<td>43</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>Total</td>
<td>329</td>
<td>361</td>
</tr>
</tbody>
</table>

Table 4. Answers to: “Is the head split?” with altered marginal values

Agreement according to the benchmarks of Landis and Koch turned to substantial with a kappa of 0.80. Absolute agreement was even slightly lower (0.90 vs. 0.95) (Table 5).

<table>
<thead>
<tr>
<th>2D - n = 46</th>
<th>Question</th>
<th>k</th>
<th>Classification</th>
<th>Percentage of overall agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Study</td>
<td>Is the head split?</td>
<td>0.26</td>
<td>Fair</td>
<td>0.95</td>
</tr>
<tr>
<td>Hypothetical Study</td>
<td>Is the head split?</td>
<td>0.80</td>
<td>Substantial</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 5. Kappa values vs. percentages of overall agreement in original study and hypothetical study

In the original study, question four was: “Is the glenohumeral joint dislocated”? Agreement of this question was ‘slight’ with a kappa of 0.12; also a counterintuitive finding considering the absolute agreement of 0.87 (Table 1).
Agreement in this study was 0.87 (Table 1). We again turned towards the case distribution; none of the cases had a dislocated glenohumeral joint; producing a substantial imbalance in ‘yes’ and ‘no’ answers. Altering the results like question 2 would lead to the same substantial agreement.

**DISCUSSION**

Agreement between observers is quantified according to the kappa coefficient by Cohen and categorized by Landis and Koch. The advantage of the kappa coefficient is its correction for the amount of agreement that can be expected to occur by chance alone. This distinct feature of the kappa coefficients has made it the instrument of choice in observer agreement studies\(^2\,8,\,13,\,19-21,\,29,\,30\). However, many investigators are not aware that kappa is affected by the case and answer distributions of their study.

Because of this feature, two observers who have high absolute agreement may generate low kappa coefficients. Cichetti and Feinstein provided us with a reasons for these contradictory low kappa coefficients: Prevalence of the subject under study in the data set has an effect on marginal values –i.e. balance or symmetry of ‘yes’ versus ‘no’ answers - as we have demonstrated in the hypothetical study. The phenomenon that a kappa coefficient may be low in contrast to high actual percentage of agreement has been dubbed the Kappa Paradox\(^5,\,10\).

Some authors point out that the Kappa Paradox is not a weakness in the kappa statistic, but rather a logical consequence of its purpose; to correctly interpret the agreement in a study is to correct this for the expected amount of agreement by chance. And if this chance is affected by the way cases and answers are distributed, it is only logical that the kappa statistic does as well\(^26\).

As for an example in daily practice; the presence of scaphoid fracture is an uncommon finding on plain radiographs, and would therefore subsequently render low kappa values of agreement if a study would be designed mimicking this distribution (e.g. 5% of plain radiographs with scaphoid fractures and 95% without)\(^1\). If one would decide to design an observer reliability study of radiographs for the presence or absence of a scaphoid fracture and would balance the number of cases with a fracture and without a scaphoid fracture the Kappa coefficients would improve.

As illustrated in Table 1, the absolute percentage of overall agreement in our published study\(^1\) was much higher than agreement according to the kappa coefficient. We contribute this finding to the Kappa Paradox as illustrated by rearranging case and answer ratios, even though this hypothetical prevalence after re-distributing cases does not represent prevalence in daily practice.

We would advise investigators to balance their case distribution to ensure that the amount of agreement that can be expected by chance is as low as possible, thus ensuring that the kappa coefficient nears the percentage of agreement.

When interpreting the results interobserver studies to date, it is vital to take the study design into account; especially with low kappa’s one should look to the case distribution before accepting the low agreement as a fact and drawing any consequences from the result.
REFERENCE LIST


