Eur J Med Res (2006) 11: 119-122 © I. Holzapfel Publishers 2006

The Comparison of Clinical Imaging Devices with Respect to Parallel Readings in Both Devices

F. Krummenauer

Clinical Epidemiology and Health Economy Unit (Head of the Unit: Prof. Dr. F. Krummenauer), University Hospital Carl Gustav Carus, Dresden University of Technology, Dresden, Germany

Abstract

Objective: Many proposals for the comparison of diagnostic devices refer to the computation of ROC curves or sensitivity / specificity-based parameters, thereby strictly assuming the presence of a reliably parametrized clinical reference method. When none of the devices under consideration can be regarded as a reference, Cohen's kappa coefficient for assessing the methods' relative agreement becomes increasingly popular. If, however, not only the agreement between two diagnostic devices, but also the devices' reliability must be taken into account (for example, if multiple parallel readings are obtained from one or both of the devices), no corresponding coefficients can be obtained from standard software. Bearing the recent modifications in the German Medicinal Devices Law (Medizinproduktegesetz) in mind, such methods will soon become necessary and strongly demanded for the sake of immediate re-evaluation of previously certified medicinal devices.

Methods: Generalizations of Cohen's kappa (κ) for complex multi reader designs can be found by estimating weighted averages of the observed and expected agreement among subsets of parallel readings. A flexible, although instructive, strategy for designing kappa coefficients in the context of method comparison trials is proposed, which measures the two methods' overall agreement while correcting for each method's underlying inter / intra observer reliability. Cluster algorithms will be outlined, which allow to identify (in)compatible clusters of readings. Their application will be illustrated by means of the intraindividual comparison of two different strategies in radiographical imaging, where none of the underlying imaging methods can be regarded as a reference.

Results: The algorithms are illustrated by the comparison of two radiological imaging devices R and F, where none of these imaging methods could be considered as a valid reference, i.e. replicate readings by three independent radiologists were taken from each device, respectively. The setting allowed for intraindividual comparison of the imaging methods, since each of the three involved radiologists took one reading from both devices on each of 120 individuals. The algorithm identifies a subset of compatible reading patterns with an overall agreement of $\kappa = 0.83$ (95% confidence interval 0.78 – 0.88) despite the fact, that the underlying readings arose from two different imaging devices. An obvious interpretation suggests, that the gradient in experience between the readers was more relevant to their reading patterns' outcome than any difference between the imaging devices.

Conclusions: The generalized κ coefficients can be modified according to the study design at hand to instructively identify (in)compatible clusters of multiple parallel reading patterns; the relative agreement of imaging methods can be estimated as well as each imaging method's internal reliability as assessed by parallel readings from the respective methods.

Key words: Cohen's kappa, agreement analysis, parallel readings, clinical imaging

1. Introduction

The need for minimum invasive, but still maximum valid and reliable diagnostic procedures encouraged the development of clinical imaging methods through the past decades. As soon as an imaging procedure was proven to show sufficient diagnostic quality, its integration into routine procedures could be enforced. Meanwhile a lot of such clinical imaging strategies are available and are still under further development concerning, for example, cost and time effectiveness.

However, introduction of an improvement of established imaging procedures into clinical routine affords an appropriate evaluation of its diagnostic potential as well as the assurance of sufficient agreement with a di-

*This paper is sincerely dedicated to my former teacher and mentor Prof. Dr. *Jörg Michaelis* (president of Johannes Gutenberg University of Mainz and former head of the department of Medical Biometry, Epidemiology and Informatics hosted at the University of Mainz) for the occasion of his 65th anniversary 07.12.2005.

The paper was presented during the 2004 annual meeting of the German Region of the *International Biometrical Society* (IBS-DR), March 2004 in Heidelberg, and during the 2005 annual meeting of the *German Society of Orthopedics and Orthopedic Surgery* (DGOOC), September 2005 in Berlin.

Financial Interests: The author has no commercial or political interests in the methodological or medical aspects presented in this paper.

agnostic reference. If both the diagnostic novum and the reference are already parameterized to derive binary findings (e.g. "positive" versus "negative") from the imaging procedures' output, a standard strategy of method comparison is the computation of sensitivity, specificity and predictive values. If at least the reference provides binary results, an additional ROC analysis can be applied instead to evaluate the novum's diagnostic potential.

Cohen's kappa coefficient (κ), however, has been established for those method comparison settings, which derive categorical findings from both imaging procedures, but where none of these procedures can be regarded as a fully valid and reliable standard. Relative agreement between two imaging methods is then estimated by comparison of the observed order of reader agreement with the corresponding order of agreement, which would have been expected by chance. A lot of generalizations of Cohen's original proposal for the comparison of two binary parallel readings have then been proposed (Landis and Koch 1977, Davies and Fleiss 1982, Dunn 1992, Kraemer 1992, Shoukri 2003) to provide kappa point and interval estimates for more complex multi reader designs. Whereas these strategies can be applied to several parallel readings in one imaging method to estimate its imaging reliability, these approaches can hardly be extended to the problem of imaging method comparison, when both imaging methods involve parallel readings.

To achieve such a multi-reader kappa estimate for method comparison, which estimates the imaging methods' relative agreement after adjustment of the respective methods' inter / intra reader reliability, a more flexible approach to estimator construction seems necessary. Bearing the recent modifications in the German Medicinal Devices Law (Medizinproduktegesetz) in mind, such methods will soon become necessary and strongly demanded for the sake of immediate re-evaluation of previously certified medicinal devices.

One of the most promising approaches is based on the appropriate pairwise comparison of readers (Schouten 1980, 1982). Instead of averaging pairwise kappa estimates, which might introduce severe bias (Dunn 1992), this paper seeks to point out an application of Schouten's approach to the method comparison of clinical imaging devices, when both devices afford multiple parallel readings and therefore consideration of intra and inter device reliability (where the intra reliability will usually be determined by the agreement between parallel clinical readers).

2. Methods

2.1 Kappa Coefficients

For a formal parameterization of kappa and its estimators the reader is refered to Schouten's model (Schouten 1980, 1982). Unformally, if two readers are compared along a nominal binary outcome variable, and o and e denote their observed and randomly espected agreement, kappa can be estimated as $\kappa = (o - e) / (1 - e)$. The observed agreement term o descibes the order of concordant findings between the parallel reading patterns, whereas the randomly expected agreement term e describes the order of agreement between two parallel clinical readers, which must already be expected by chance. The above κ coefficent then measures the order of "competence" based agreement between two parallel readers, i.e. the order of agreement not only due to randomly concordant findings. $\kappa = 0$ indicates only random agreement (since $o = e$), the larger κ becomes, the more evidence for inter reader agreement and thereby diagnostic reliability can be established.

If $r > 2$ parallel observers are asked to derive binary readings ("negativ" or "positive" finding) on a subset of individuals, these observers' pairwise observed and expected agreement can be estimated accordingly; the estimates shall be denoted $o^{(a,b)}$ and $e^{(a,b)}$ for two observers a and b. The overall kappa coefficient for this team of parallel readers is then estimated by first averaging the pairwise observed agreement estimates $o^{(a,b)}$, then the pairwise expected estimates $e^{(a,b)}$, and then introducing the resulting averaged estimates o and e into the above kappa expression. Asymptotic interval estimation now becomes feasible by strict application of the delta method (Schouten 1980, 1982). Details for this estimation strategy in the setting under consideration are presented by Krummenauer (Krummenauer, 2005).

2.2 Cluster Algorithms for Multi-reader Designs

It is straight-forward to apply the idea of averaging kappa ingredients to the previously outlined setting of method comparison under adjustment for multi reader reliability in both imaging devices under consideration. If r_1 parallel readings are taken from imaging method 1 and r_2 from method 2, then the total set of $r_1 + r_2$ parallel readings can be introduced into a stepwise cluster algorithm to identify incompatible reading patterns: In a first step an overall agreement of the $r_1 + r_2$ readings / readers is estimated according to the above pairwise averaging procedure; then each of the $r_1 + r_2$ readers is contrasted to the remaining ones, respectively, and a pairwisely averaged kappa is estimated by pairwise comparison of all $r_1 + r_2 - 1$ readers with the index reader. If the resulting overall agreement estimate and the contrasting one differ significantly, an indication for removal of this index rater from the overall team is found.

A sequential algorithm would then eliminate this one reader, whose contrasting agreement kappa shows maximum deviation from the overall kappa: It is easy to show, that the agreement of the remaining team of $r_1 + r_2 - 1$ readers shows larger overall agreement than the previous team, as soon as the contrasting agreement kappa turns out smaller than the overall agreement kappa. This eliminating strategy can be based on the above a significance test for kappa estimator comparison and therefore presents a basis for a step-down cluster algorithm (Krummenauer 2005). This algorithm will reduce the reader team, as long as there are readers with statistically significant deviation between the actual overall and their actual contrasting agree-

ment κ estimate. If the algorithms stops, the resulting reading patterns can be regarded as no longer significantly incompatible. If the remaining readings all emerged from the same imaging method and the readings based on the other were eliminated, an exploratory interpretation of method disagreement is at hand.

A corresponding step-up analogue of this procedure would start with pairwise comparison of readers and "cluster" those readers, whose reading patterns show maximum inter-reader agreement. Of course, this algorithms is of an even more exploratory nature than the previous step-down formulation.

3. Example

The step-up and step-down algorithms based on pairwisely averaging kappa estimators will be briefly illustrated by the comparison of two radiological imaging devices R and F. The example is based on real data. However, since the underlying clinical data is unpublished yet, the imaging devices will not be further specified,; the data will only be used for the sake of illustration.

None of the imaging methods could be considered as an error-free reference, i.e. replicate readings by three independent radiologists were taken from each device, respectively. The setting allowed for intraindividual comparison of the imaging methods, since each of the three involved radiologists took one reading from both devices on each of 120 individuals. Reader 1 had a 15 years experience in clinical imaging, readers 2 and 3 could be considered quite less experienced.

Table 1 presents the results of the step-down eliminating algorithm, which first (at a local significance level 5%) deleted the readings of reader 1 based on device R, and then on device F. Afterwards the algorithm stopped with a remaining set of reading patterns, which included two parallel readings from device R and F, respectively; these four readings did not differ significantly in terms of the kappa comparison outlined above. The remaining subset of readings shows a reliability of $\kappa = 0.83$ (asymptotic 95% confidence interval 0.78 – 0.88) – despite the fact, that the underlying readings arose from two different imaging devices!

An obvious interpretation suggests, that the gradient in experience between reader 1 and the others was more relevant to their reading patterns' outcome than any difference between the imaging devices. Table 2 confirms this impression by presenting interim results of the step-up algorithm, where clusters were aggregated, when the kappa point estimate between them was larger than 0.50: After two clustering steps the F and R readings of reader 1 gather with an inter-device kappa estimate of 0.81 (0.76 – 0.87), whereas the "inter-cluster" kappas between the other reading patterns suggest aggregation of reading clusters R_2, F_3 first with F_2 and then with R_3. The algorithm stopped with the clusters R_1, F_1 versus R_2, R_3, F_2, F_3 and respective "intra-cluster" kappa estimates of 0.81 (0.75 – 0.86) and 0.75 (0.71 – 0.79). These clusters show an "inter-cluster" kappa estimate of 0.32 (0.28 – 0.37) and therefore were not aggregated in a further step.

Table 2. Interim results of a step-up cluster algorithm to identify compatible reading patterns among six parallel readings on 120 subjects: inter-team kappa estimates for the respective subset of 6 parallel readings; $R_1 – R_3$ and $F_1 – F_3$ denote the readings based on imaging devices R and F by three independent clinical readers, respectively.

	R_1, F_1	**R_2, F_3**	**R_3**	**F_2**
R_1, F_1	0.81	0.32	0.23	0.35
R_2, F_3		0.78	0.62	0.69
R_3				0.56

4. Discussion

The pairwise agreement concept of Schouten (1980, 1982) was applied to implement step-up and step-down selection algorithms for an exploratory comparison of two clinical imaging devices, when both show restrictions in reliability and therefore call for parallel readings. The algorithms gather or eliminate single readings from an overall reading pattern of multiple parallel readings by local significance tests (step-down) and by descriptive interpretation of point estimates (step-up).

Table 1. Results of a step-down cluster algorithm to identify compatible reading patterns among six parallel readings on 120 subjects: kappa estimates for the complete set of readings (last column) and inter-team kappa estimates for each reading versus the remaining ones (columns 2 – 7), p values derived from asymptotic tests for the comparison of the total kappa estimate and the respective inter-team estimates; $R_1 – R_3$ and $F_1 – F_3$ denote the readings based on imaging devices R and F by three independent clinical readers, respectively.

	R_1	**R_2**	**R_3**	**F_3**	**F_2**	**F_1**	**total κ**
κ (%)	51	66	66	71	64	59	64
p	0.01	0.42	0.42	0.16	0.90	0.19	
κ (%)		72	67	73	66	68	72
p		0.61	0.55	0.57	0.20	<0.01	
κ (%)		85	79	85	80		83
p		0.90	0.41	0.85	0.37		

The above proposals can surely be improved in several directions: First it is not only important, whether single readings differ significantly from the others (i.e. tests on differences are used); it would be desirable to decide about elimination of raters by means of one-sided equivalence tests (i.e. not any loss, but rather a clinically relevant loss in agreement could be used as a rationale for reading comparisons). In the same sense, the step-up algorithm should not only be based on the exploratory interpretation of "inter-cluster" point estimates, but rather on tests for clinically relevant increase in kappa estimates by aggregation of clusters (note that cluster condensation will become quite liberal by means of the above approach).

It must be emphasized that any kappa based approach to the analysis of multi reader data is only an exploratory, but not a modelling one (Becker and Agresti, 1991): The results of the example in section 3 can only be regarded as results obtained on a test data set, i.e. result validation on an independent data set is still necessary. Furthermore the interpretation of analyses such as described in section 3 is only based on the results of optimization along the remaining reading patterns' overall agreement. It is possible, that the most experienced reader, whose findings are most valid, becomes eliminated first, since his results differ from those of the other readers just because of his higher competence in clinical reading. This could be an explanation of the separation result in section 3, where the most experienced reader's findings from both devices are more compatible among each other than with readings from the respective devices. Therefore the optimization of reliability as achieved by the cluster algorithms can imply a notable loss in validity of the remaining readings. Unfortunately, there was no clinical "external" reference available for the imaging data in section 3; i.e. the "experience" conclusion mentioned there could not be confirmed by introduction of a reference reading as a 7th parallel reading into the step-down algorithm.

Conclusion

The kappa estimators according to Schouten's pairwisely averaging construction principle can instructively identify (in)compatible clusters among multiple parallel readings; the relative agreement of imaging methods can be estimated as well as each imaging method's respective reliability as characterized by parallel reading from the underlying imaging device. However, interpretation of these estimates must sensitively consider the underlying nature of the data. Several possibilities to improve the selection algorithms strongly motivate further method comparison research based on this approach.

Acknowledgement: The author is grateful to Ms Karen Faulkner (medical student) for a native speaker revision of this manuscript.

References

Becker MP, Agresti A. Loglinear modeling of pairwise interobserver agreement on a categorical scale. Statistics in Medicine 1991; 10: 101-14

Davies M, Fleiss JL. Measuring agreement for multinomial data. Biometrics 1982; 38: 1047-51

Dunn G. Design and analysis of reliability studies. 1992, Edward Arnold, London

Kraemer HC. Measurement of reliability for categorical data in medical research. Statistical Methods in Medical Research 1992; 1: 183-99

Krummenauer F: Methoden zur Evaluation bildgebender Verfahren von begrenzter Reproduzierbarkeit. 2005, Shaker Verlag, Aachen

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-74

Schouten HJA. Measuring pairwise agreement among many observers. Biometrical Journal 1980; 22: 497-504

Schouten HJA. Measuring pairwise agreement among many observers: some improvements and additions. Biometrical Journal 1980; 24: 318-22

Shoukri MM. Measures of interobserver agreement. 2003, Chapman & Hall, London

Received: December 9, 2005 / Accepted: February 15, 2006

Address for correspondence:
Prof. Dr. Frank Krummenauer
Clinical Epidemiology and Health Economy Unit
Dresden University of Technology
Fetscherstr. 74; Haus 29
D-01307 Dresden (Germany)
Phone: +49-351-458 3747
Fax: +49-351-458 4344
Email: Frank.Krummenauer@uniklinikum-dresden.de