http://jvme.utpjournals.press/doi/pdf/10.3138/jvme.1111.113R - Sunday, September 17, 2017 11:36:48 PM - Ecole Nationale Veterinaire D'Alfort IP Address:194.254.35.128

Inter-observer Agreement on a Checklist to Evaluate Scientific Publications in the Field of Animal Reproduction

Céline Simoneit ■ Wolfgang Heuwieser ■ Sebastian P. Arlt

ABSTRACT

This study's objective was to determine respondents' inter-observer agreement on a detailed checklist to evaluate three exemplars (one case report, one randomized controlled study without blinding, and one blinded, randomized controlled study) of the scientific literature in the field of bovine reproduction. Fourteen international scientists in the field of animal reproduction were provided with the three articles, three copies of the checklist, and a supplementary explanation. Overall, 13 responded to more than 90% of the items. Overall repeatability between respondents using Fleiss's κ was 0.35 (fair agreement). Combining the "strongly agree" and "agree" responses and the "strongly disagree" and "disagree" responses increased κ to 0.49 (moderate agreement). Evaluation of information given in the three articles on housing of the animals (35% identical answers) and preconditions or pretreatments (42%) varied widely. Even though the overall repeatability was fair, repeatability concerning the important categories was high (e.g., level of agreement = 98%). Our data show that the checklist is a reasonable and practical supporting tool to assess the quality of publications. Therefore, it may be used in teaching and practicing evidence-based veterinary medicine. It can support training in systematic and critical appraisal of information and in clinical decision making.

Key words: evidence-based veterinary medicine, literature quality, evaluation form

INTRODUCTION

It is fundamental that veterinarians use the most current diagnostic and therapeutic interventions with the best available evidence to achieve the best care for their patients. *Evidence* can be defined as the extent of sureness that scientific findings are true.[1] The conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients is called *evidence-based medicine* (EBM).[2] In the biological sciences, including current veterinary medicine, the pool of available information increases exponentially.[3–5] The scientific literature is considered one of the most influential information sources in medicine,[6] and it is an important link between research and practice.[7] Consequently, to stay up to date the veterinarian has to select and assimilate an enormous amount of information. In addition, publications in veterinary journals vary widely in their quality.[8,9] Therefore, it is essential to train students and practitioners in how to retrieve information and assess its quality.

In human medicine, levels of evidence have been ranked from the strongest to the weakest[10] and applied to the hierarchic staircase of evidence (Figure 1).[11] Ideally, clinical decision making should be founded on the highest level of evidence available for the specific question.[12] Randomized controlled trials (RCTs) are the gold standard to achieve a high level of evidence. Lower evidence levels do not inevitably imply that results or conclusions are wrong. The reader, however, has to consider that the certainty with which the results represent the truth is weaker. The hierarchy-of-evidence concept further recognizes that RCTs are not suitable for answering all types of clinical questions; for example, RCTs are not suitable for estimating the influence of season on the reproductive performance of cattle because many parameters of a control group, such as feeding and air temperature, would distort and conceal seasonal effects. However, flaws that relate to design, execution, and reporting represent a threat to the validity of study findings.[12] Hence, aside from determining the level of evidence, it is important to identify specific deficits of trials and publications to determine their quality and critically evaluate the validity and practicability of their findings. For students and practitioners not experienced in handling scientific information, evaluating validity and suitability for the management of a given case is a challenge.[13] To alleviate this problem, the use of checklists has been suggested.[8] Several authors have reported the use of such tools in evaluating scientific publications in human medicine. For the assessment of veterinary literature, few detailed checklists have been published.[8,9] A brief checklist was designed to assist veterinary students in appraising the quality of literature.[14] Its authors concluded that the checklist was helpful in evaluating the publications and revealing deficits. The different quality aspects of information and the use of checklists for an objective and effective evaluation should be integrated into veterinary education.

In the broadest sense, a checklist to evaluate the quality of scientific literature can be seen as a diagnostic tool with which independent reviewers should be able to obtain similar results.[15] Therefore, high agreement between students or veterinarians using a checklist is important to demonstrate its validity and usability. The objective of this study was to determine inter-observer agreement

doi:10.3138/jvme.1111.113R

http://jvme.utpjournals.press/doi/pdf/10.3138/jvme.1111.113R - Sunday, September 17, 2017 11:36:48 PM - Ecole Nationale Veterinaire D'Alfort IP Address:194.254.35.128

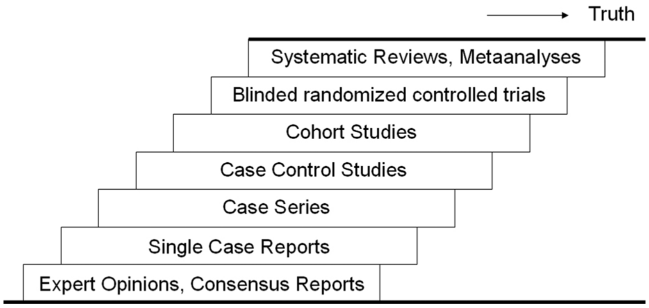
Truth
Systematic Reviews, Metaanalyses
Blinded randomized controlled trials
Cohort Studies
Case Control Studies
Case Series
Single Case Reports
Expert Opinions, Consensus Reports

Figure 1: Staircase of evidence

using a detailed checklist to evaluate representative scientific literature in the field of bovine reproduction. The checklist was developed to be applicable to case reports as well as RCTs. Specifically, we tested (1) whether different publications were evaluated with similar results by independent observers, (2) whether inter-observer agreement differs for specific criteria of the checklist, and (3) whether the observers classify each publication as the same level of evidence. In addition, we wanted to test whether the checklist was intuitively clear enough to allow reviewers to apply it without prior training.

MATERIAL AND METHODS

We selected three publications on bovine reproduction. Two were published in peer-reviewed journals (one RCT without blinding[16] and one blinded RCT [BRCT]),[17] and one was published in a journal that was not peer reviewed (a case report [CR]).[18]

To evaluate the literature, the checklist (Appendix 1), which had recently been developed,[9,19] was used. It consists of 40 criteria in these categories: materials and methodology, study design, statistics, presentation and information content, applicability, and conclusions. During the development of the checklist, the Consolidated Standards of Reporting Trials (CONSORT) Statement[20] and other publications on EBM[21–23] were taken into account. The observer was able to indicate the level of agreement with each criterion on a five-point Likert scale: *strongly agree*, *agree*, *neutral*, *disagree*, or *strongly disagree*.[24] Statements that were not accessible or applicable were characterized as "not determined." The level of evidence could be classified as "meta-analysis," "randomized controlled trial," "controlled trial without randomization," "not experimental descriptive trials," "case report," or "personal experience or expert opinion."

We asked 30 scientists in the field of animal reproduction affiliated with universities in Austria, Belgium, Germany, and Switzerland to take part in the study. Each reviewer was provided with the three articles, three checklists, and supplementary explanations. The latter gave definitions of relevant technical terms. One reviewer who had published research papers on EBM and taught courses in EBM to veterinary students for more than five years had also participated in the development of the checklist. All other reviewers had limited or no prior documented experience in the field of EBM. Study participants were informed of the identities of the other reviewers.

Data were analyzed using SPSS for Windows[a] using an additional syntax for κ.[b] Answers were coded as 1 (strongly agree), 2 (agree), 3 (neutral), 4 (disagree), 5 (strongly disagree), and 6 (not determined). When the responses "strongly agree" and "agree" and "strongly disagree" and "disagree" were aggregated, responses were coded as 1 through 4, respectively. We used Fleiss's κ test to estimate the inter-observer agreement of the 13 scientists for all three publications, each publication by itself, and groups of criteria. The possible κ values range from -1 to 1, although values usually fall between 0 and 1.[25] Values approximating zero were interpreted as close-to-chance agreement, as though the evaluator had simply guessed on every rating. Values less than zero were interpreted as worse-than-chance agreement.[26] Landis and Koch[27] interpreted κ values as follows: less than 0.00 = poor, 0.00–0.40 = slight to fair, 0.41–0.80 = moderate to substantial, and more than 0.80 = almost perfect agreement.

RESULTS

Although 21 of the 30 (70%) scientists contacted initially indicated interest in participating in the study, only 14 (47%) returned the evaluation forms within three months. In total, 42 evaluations forms were completed. Overall, 13 responded to more than 90% of the statements on the three evaluation forms (100% of the items were rated by four respondents, 99.2% by two, 98.3% by two, 97.5% by two, 96.7% by one, 94.2% by one, and 91.7% by one) and one responded to 69.2% of the statements. The most frequent unanswered statements were "Description of material is clear and detailed," "The examinations are described in detail," and "The study design is described in detail regarding prospectivity and retrospectivity." The majority of missing data concerned the BRCT (RCT, 15.7%; BRCT, 52.9%; and CR, 31.4%).

The κ test is one of the most common ways used to assess agreement on categorical and continuous variables.[28] Fleiss's κ is only calculated for criteria that are evaluated by all respondents. Hence, criteria that were not evaluated by all reviewers were not included in the tests. Overall, 92 of 120 criteria (77%) were evaluated by all reviewers. For repeatability of the reviewers evaluating all three articles, $\kappa = 0.35$. Combining the responses "strongly agree" and "agree" and "strongly disagree" and "disagree" increased the κ value to 0.49.

Some selected responses of the 13 reviewers are given in Tables 1–3. Very strong agreement was found concerning the classification of the level of evidence, which was based on the correct definition of all but one article (41 of 42 articles; 97.6%). One reviewer considered a RCT a controlled trial. Moreover, high agreement of 90% and 80%, respectively, was found regarding the assessment of a controlled study design and randomization of the trials. Total agreement was found regarding the criteria of the case report (100%), such as study design (consisting of a control group, randomization, or blinding) and questions regarding the statistics. Information provided concerning housing and preconditions or pretreatments of the animals did show variable estimations for all three publications. Of the reviewers, 35% and 42%, respectively chose the same answer concerning the criteria of housing

http://jvme.utpjournals.press/doi/pdf/10.3138/jvme.1111.113R - Sunday, September 17, 2017 11:36:48 PM - Ecole Nationale Veterinaire D'Alfort IP Address:194.254.35.128

Table 1: Selected answers of the evaluation of a randomized controlled trial of 13 reviewers using a checklist

Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not determined	Missing
Description of material is clear and detailed.	3	8	0	0	0	0	2
Information about treatments or interventions is adequate.	10	3	0	0	0	0	0
The study was controlled.	13	0	0	0	0	0	0
The study was randomized.	12	0	0	0	0	1	0
The study was blinded.	1	0	0	0	5	6	1
Description of study design is clear and detailed.	6	7	0	0	0	0	0
Data are sufficient to draw valid conclusions.	1	8	2	1	0	0	1

Table 2: Selected answers of the evaluation of a blinded randomized controlled trial of 13 reviewers using a checklist

Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not determined	Missing
Description of material is clear and detailed.	2	6	4	0	0	0	1
Information about treatments or interventions is adequate.	7	4	1	0	0	0	0
The study was controlled.	13	0	0	0	0	0	0
The study was randomized.	13	0	0	0	0	0	0
The study was blinded.	12	0	0	0	1	0	0
Description of study design is clear and detailed.	3	4	5	0	0	0	1
Data are sufficient to draw valid conclusions.	0	10	2	1	0	0	0

Table 3: Selected answers of the evaluation of a case report of 13 reviewers using a checklist

Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not determined	Missing
Description of material is clear and detailed.	1	6	5	0	0	0	1
Information about treatments or interventions is adequate	0	2	1	0	0	10	0
The study was controlled.	0	0	0	0	4	9	0
The study was randomized.	0	0	0	0	4	9	0
The study was blinded.	0	0	0	0	4	9	0
Description of study design is clear and detailed.	1	1	0	0	0	11	1
Data are sufficient to draw valid conclusions.	0	3	4	4	2	0	0

and preconditions or pretreatments. Concerning the RCT, responses regarding information about the age of the animals (40% identical answers) and the description of the study design regarding the type of blinding (36%) and the handling of missing data (31%) showed high variability. Assessment of the critical discussion of results (36%), the applicability of the data (36%), and the adequacy of the data to draw valid conclusions (31%) for the case report were also highly variable. The κ values for the study design ($\kappa = 0.64$) and statistics ($\kappa = 0.63$) categories demonstrated much better repeatability among reviewers than values concerning materials and methodology ($\kappa = 0.36$) or presentation and applicability ($\kappa = 0.28$) (Table 4). Higher agreement was found ($\kappa = 0.49$) for the case report than for the RCT without blinding ($\kappa = 0.13$) and the BRCT ($\kappa = 0.29$) (Table 5).

DISCUSSION

Training in evidence-based veterinary medicine should encompass a thorough exercise in critically appraising the quality of scientific information.[14] An applicable tool for students and for practitioners would be helpful in systematically assessing specific quality parameters. In courses, it could also be used as an example in discussing the specific quality aspects of evidence levels and additional criteria. The objective of this trial was to evaluate inter-observer agreement in using the detailed checklist, which can be used for case reports as well as RCTs on animal reproduction. In addition, we wanted to test whether the checklist was intuitively clear enough that reviewers could apply it without prior training. Hence, the reviewers intentionally did not receive any formal training in the application of the checklist. Furthermore,

http://jvme.utpjournals.press/doi/pdf/10.3138/jvme.1111.113R - Sunday, September 17, 2017 11:36:48 PM - Ecole Nationale Veterinaire D'Alfort IP Address:194.254.35.128

Table 4: Kappa values and confidence limits of clustered answers of three publications assessed by 13 reviewers using a checklist

Criteria	Kappa	Confidence limit
Study design	0.64	0.57, 0.71
Statistics	0.63	0.56, 0.70
Materials and methodology	0.36	0.26, 0.47
Presentation and applicability	0.28	0.18, 0.39
Overall	0.49	0.45, 0.54

Table 5: Kappa values and confidence limits of evidence levels assessed by 13 reviewers using a checklist

Level of evidence	Kappa	Confidence limit
Case report	0.49	0.43, 0.56
Randomized controlled study without blinding	0.13	– 0.08, 0.34
Blinded, randomized controlled trial	0.29	0.19, 0.39

reviewers could not discuss or compare their ratings. Nevertheless, an additional form with definitions was provided to explain technical terms used in the checklist.

The selection of the scientists could have led to bias. Selecting different reviewers might have led to different results. Reasons could be, for example, diverse involvement in research projects or different expectations of the quality of scientific information. However, it was important to choose reviewers who were experts in the field of the content of the articles to avoid bias caused by different educational status. Whether practitioners and students are also able to use the checklist with reliable results remains open to question. Therefore, we suggest evaluating whether the checklist is a helpful tool for these target groups as well.

In this study, we selected three articles of different quality to include relevant evidence levels typical of the current available literature in veterinary journals.[8–10] Clinical trials lacking randomization or blinding and case reports represent most of the actual available literature in veterinary medicine.[9] Only a few meta-analyses have been published on veterinary topics.[29] The three articles selected may have led to bias. Whether the selection of different publications would have led to a different inter-observer agreement remains open. Nevertheless, we had to select a sample that represented different evidence levels and was able to be evaluated in reasonable time.

Most of the respondents filled out the evaluation forms completely. They were asked to give additional comments regarding the evaluation and usability of the checklist. However, this option was not used. Some statements may have been left unanswered because a statement was unclear or the reviewer was unable to find a valid answer. Because the reviewers were specialized in the field of animal reproduction, whether they had a working knowledge of study design and quality remains speculative. They may have considered some questions unimportant or overlooked them. Another issue might have been a lack of time. One respondent did not fill out a whole page of the checklist. The case report is the publication with the most missing data (RCT, 23.9%; BRCT, 28.3%; CR, 47.8%). Overall, there was high variability between reviewers in the criteria left blank. Maybe some of the criteria should be reformulated for easier understanding. Reviewers did not provide comments to support this hypothesis; hence, we cannot formulate a sound conclusion as to whether the checklist was sufficiently intuitively clear to allow reviewers to apply it without prior training.

The checklist has recently been used in different literature evaluation projects. In addition, other institutions may use it in evaluating existing literature or for training purposes. We are grateful for any comments to improve the checklist and its use.

Likert scales are commonly used in checklists[30] but may be subject to distortion. One reason is that the Likert scale measures both directions (agree and disagree) and strengths, which can lead to an underestimation of the extreme positions, strongly agree and strongly disagree.[31] Moreover, confusion occurs with an odd number of responses.[9] The midpoint statement "neutral" may be confused with "don't know" or "not available."[32] For this reason, we provided the response "not determined." Future studies should investigate whether larger scales, a three-point scale, or a simple yes–no option provide higher reliability.

The κ test is the preferred statistical procedure to estimate inter-observer agreement between two respondents for nominal or ordinal-scale data.[26] Fleiss's κ can be adapted for more than two respondents.[33] In some publications, the intraclass correlation coefficient is given as an indicator of agreement. However, it is designed to assess consistency or conformity between two or more quantitative measurements.[34] Therefore, we could not use it for the ordinal data in this study. Responses were regarded as independent data. Nevertheless, in some cases the assessment of a criterion may have influenced the estimation of another one, which might have led to distortion of the results regarding a single criterion. Different interpretations of κ test results have been published. Nevertheless, they all have specific limitations. The scores of Landis and Koch[27] are broadly accepted and were therefore taken into account.

Generally, assignment of evidence levels is based only on the reported study design and results and not on the quality of the data or its interpretation.[15] Therefore, our checklist included additional criteria to help reviewers estimate the reliability of the data. Our results showed that the agreement between the reviewers varied with the criterion evaluated. The lower repeatability of information on materials and methodology such as housing, age, and pretreatment of the animals as well as presentation and applicability could accrue from inadequate reporting or different opinions concerning adequate reporting. Possibly, not all respondents knew specifically what information should be reported according to current standards specified by the CONSORT Group (i.e., the CONSORT Statement, which aims to improve the reporting of RCTs).[20] Regarding the case report, we found a higher agreement in total and on criteria specifying the study design and the statistical procedures. All respond-

http://jvme.utpjournals.press/doi/pdf/10.3138/jvme.1111.113R - Sunday, September 17, 2017 11:36:48 PM - Ecole Nationale Veterinaire D'Alfort IP Address:194.254.35.128

ents chose "not to be determined." In addition, some respondents chose "strongly disagree" regarding the criteria "the study was controlled," "randomization was conducted," or "the study was blinded," whereas others chose "not determined." This inconsistency in the records led to a decrease in the κ value, although the reviewers intended a similar statement. Thus, the option "not determined" seems to have caused a higher variability in the answers. Nevertheless, we found this option helpful in addressing not determinable criteria. Whether a specific checklist for each evidence level would be advantageous remains open to question. However, our aim was to develop a test instrument that is broadly feasible to use in education and practice and covers the most relevant validity aspects.

Overall, the reviewers agreed in identifying the level of evidence of a publication; 98% chose the adequate level of evidence. Only one respondent classified one article as a controlled trial, which is just one evidence level beneath the RCT level. Despite this error, the evaluator strongly agreed that randomization was conducted and agreed that the study design was described in detail regarding the type of randomization. Therefore, the use of detailed criteria besides the level of evidence seems to be relevant. The responses concerning the adequacy of the data to draw valid conclusions in the CR were highly variable (31% identical answers), although overall the CR was consistently classified as a low level of evidence. Concerning the RCTs, the respondents agreed mostly on the high quality of the information.

Overall, our findings demonstrate an agreement among the respondents. Combining the extreme-positions values raised the κ from fair ($\kappa = 0.35$) to moderate ($\kappa = 0.49$) agreement.[27] Even if reviewers did not answer some questions identically, responses tended to be in the same direction.

More recently, a similar study with eight reviewers rating 86 clinical studies on human urological literature was published. Its κ values ranged from 0.20 to 0.48, and the intraclass correlation coefficients were 0.67 regarding the type of study and 0.55 regarding the level of evidence.[12] An intraclass correlation coefficient greater than 0.75 indicates good agreement.[35] In that study, because of an initially low inter-observer agreement, levels-of-evidence subcategories (i.e., IIa, IIb, IIc) were secondarily collapsed into the main category (i.e., II). Therefore, the level of evidence as well as the type of study were categorized by four possible answers. Bhandari et al.[15] described the intraclass correlation coefficients for the agreement among six reviewers evaluating 51 clinical papers. They ranged from 0.61 (overall level of evidence) to 0.75 (type of study). Two respondents trained in epidemiology demonstrated extremely high agreement (0.99–1.0). Complete agreement regarding the level of evidence was higher in our study (98%) than in Bhandari et al.[15] (67%), and Turpen et al.[12] (12%). However, we did use only three articles that had been preselected to represent different levels of evidence, and only 47% of the participants returned the checklist. It remains open to question whether the task of classifying the level of evidence was easier in our experiment than in Bhandari et al. and Turpen et al.'s studies[15,12] (one case report and two RCTs vs. 51 and 86 clinical trials, respectively, without case reports or basic research articles).

A study carried out with veterinary students used a shorter checklist that encompassed nine specific validity aspects.[14] Using this tool, the respondent had to determine the evidence level and agree or disagree with statements given about study design, information content, and objectivity. Finally, publication rating points were summed to obtain an overall score. Of the students using the checklist, 67% assessed the correct level of evidence. Nevertheless, most of the students (82%) stated that they had considered additional criteria to evaluate the literature, compared with an assessment without a checklist.[14] The detailed checklist we present includes many more criteria for a much more detailed assessment. This may be advantageous for students and practitioners in gaining a deeper understanding of the aspects of literature quality. It could be included in courses that aim to train students in evaluating information in more detail. Therefore, we suggest testing the use of the detailed checklist in veterinary education and postgraduate education.

Studies of a high evidence level may be deemed appropriate for application to patient care, whereas lower-evidence-level studies should be interpreted with caution.[15] However, well-designed observational studies can provide results consistent with those of randomized trials.[36] In addition, meta-analysis and RCTs could just as easily report deceptive results.[37,38] Ultimately, a sound answer to a clinical question should ideally be based on a composite assessment of evidence of all types. No single study necessarily provides a definitive answer.[15] Therefore, the levels-of-evidence classification system should be viewed in the context of the clinical question, the quality of the study's methods, and the biological plausibility of the results.

Our data have shown that the detailed checklist we present, although imperfect, does provide a reasonable and practical tool to assess the publication quality. Nevertheless, as evidenced by Bhandari et al.,[15] specific training appears important for the correct assessment of study design and methodological quality. This need for training emphasizes the importance of increased educational efforts to promote the principles of evidence-based veterinary medicine.[12] A first step would be to establish the checklist as a supporting tool available to students in the first semesters. In addition, it could support assessment of information in courses that encompass case-based learning or in journal clubs.

NOTES

[a] Version 16.0; SPSS Inc., Munich, Germany

[b] Obtained from http://www.spsstools.net/Syntax/Matrix/CohensKappa.txt

ACKNOWLEDGMENTS

We thank the participating reviewers for their time and efforts. We also express our gratitude to Mr. Lotz, scientific assistant of the Department of Biometry and Statistics, Faculty of Veterinary Medicine, Berlin, for statistical support. Furthermore, we thank Ms. Dicty, scientific assistant of the animal reproduction clinic, Faculty of Veterinary Medicine, Berlin, for corresponding with the reviewers.

http://jvme.utpjournals.press/doi/pdf/10.3138/jvme.1111.113R - Sunday, September 17, 2017 11:36:48 PM - Ecole Nationale Veterinaire D'Alfort IP Address:194.254.35.128

REFERENCES

1 Arlt S, Heuwieser W. Evidenz-Basierte Veterinärmedizin [Evidence-based veterinary medicine]. Dtsch Tierarztl Wochenschr. 2005;112(4):146–8. German. Medline:15900679

2 Sackett DL, Rosenberg WM, Gray JA, et al. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71–2. http://dx.doi.org/10.1136/bmj.312.7023.71. Medline:8555924

3 Martens H. Universitäre Ausbildung heute: Notwendigkeit und Perspektiven [University education today: necessity and prospects]. Tierärztl Prax. 2001;29(K):144–1. German.

4 Baum J. Don't get left behind: three critical ways to keep your veterinary practice competitive, and profitable. In: Proceedings of the North American Veterinary Conference, vol. 2; 2008; Orlando, Florida, USA. p. 187–9.

5 Buchanan RA, Wooldridge AA. Staying current by searching the veterinary literature. J Vet Med Educ. 2011;38(1):10–5. http://dx.doi.org/10.3138/jvme.38.1.10. Medline:21805930

6 Buchanan RA, Wooldridge AA. Staying current by searching the veterinary literature. J Vet Med Educ. 2011;38(1):10–5. http://dx.doi.org/10.3138/jvme.38.1.10. Medline:21805930

7 Antes G, Bassler D. [Evidence-based medicine, dissemination of research information and the role of the medical journal]. Dtsch Med Wochenschr. 2000;125(38):1119–21. German. http://dx.doi.org/10.1055/s-2000-7574. Medline:11147365

8 Kastelic JP. Critical evaluation of scientific articles and other sources of information: an introduction to evidence-based veterinary medicine. Theriogenology. 2006;66(3):534–42. http://dx.doi.org/10.1016/j.theriogenology.2006.04.017. Medline:16720037

9 Arlt S, Dicty V, Heuwieser W. Evidence-based medicine in canine reproduction: quality of current available literature. Reprod Domest Anim. 2010;45(6):1052–8. http://dx.doi.org/10.1111/j.1439-0531.2009.01492.x. Medline:19563501

10 Schmidt PL. Evidence-based veterinary medicine: evolution, revolution, or repackaging of veterinary practice? Vet Clin North Am Small Anim Pract. 2007;37(3):409–17. http://dx.doi.org/10.1016/j.cvsm.2007.01.001. Medline:17466746

11 Arlt S, Heuwieser W. Evidence-based complementary and alternative veterinary medicine—a contradiction in terms? Berl Munch Tierarztl Wochenschr. 2010;123(9–10):377–84. Medline:21038809

12 Turpen RM, Fesperman SF, Sultan S, et al. Levels of evidence ratings in the urological literature: an assessment of inter-observer agreement. BJU Int. 2010;105(5):602–6. http://dx.doi.org/10.1111/j.1464-410X.2009.09181.x. Medline:20089109

13 Glasziou P, Guyatt GH, Dans AL, et al. Applying the results of trials and systematic reviews to individual patients. ACP J Club. 1998;129(3):A15–6. Medline:9825009

14 Arlt SP, Heuwieser W. Training students to appraise the quality of scientific literature. J Vet Med Educ. 2011;38(2):135–40. http://dx.doi.org/10.3138/jvme.38.2.135. Medline:22023921

15 Bhandari M, Swiontkowski MF, Einhorn TA, et al. Inter-observer agreement in the application of levels of evidence to scientific papers in the American volume of the Journal of Bone and Joint Surgery. J Bone Joint Surg Am. 2004;86-A(8):1717–20. Medline:15292420

16 Kasimanickam R, Duffield TF, Foster RA, et al. The effect of a single administration of cephapirin or cloprostenol on the reproductive performance of dairy cows with subclinical endometritis. Theriogenology. 2005;63(3):818–30. http://dx.doi.org/10.1016/j.theriogenology.2004.05.002. Medline:15629800

17 Friton GM, Cajal C, Ramirez-Romero R. Long-term effects of meloxicam in the treatment of respiratory disease in fattening cattle. Vet Rec. 2005;156(25):809–11. Medline:15965005

18 Arlt S, Münnich A, Schröder M. Penisverkrümmung bei einem Fleckvieh-Bullen [Deviation of the penis of a polled bull: a case report]. Prakt Tierarzt. 2005;86(11):842–4. German.

19 Simoneit C, Heuwieser W, Arlt S. Evidence-based medicine in bovine, equine and canine reproduction: quality of current literature. Theriogenology. 2011;76(6):1042–50. http://dx.doi.org/10.1016/j.theriogenology.2011.05.007. Medline:21719082

20 Consolidated Standards of Reporting Trials (CONSORT) Group. CONSORT 2010 statement [Internet]; 2010 [cited 2012 Apr 25]. Available from: http://www.consort-statement.org/consort-statement/overview0/.

21 Guyatt GH, Sackett DL, Cook DJ, et al. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA. 1994;271(1):59–63. http://dx.doi.org/10.1001/jama.1994.03510250075039. Medline:8258890

22 Kastelic JP. Critical evaluation of scientific articles and other sources of information: an introduction to evidence-based veterinary medicine. Theriogenology. 2006;66(3):534–42. http://dx.doi.org/10.1016/j.theriogenology.2006.04.017. Medline:16720037

23 Kunz R, Ollenschläger G, Raspe H, et al. Lehrbuch evidenzbasierte Medizin in Klinik und Praxis [Textbook of evidence-based medicine]. Köln: Dt. Ärzteverlag; 2000. p. 432. German.

24 Likert R. A technique for the measurement of attitudes. Arch Psychol. 1932;140:1–55.

25 Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005;85(3):257–68. Medline:15733050

http://jvme.utpjournals.press/doi/pdf/10.3138/jvme.1111.113R - Sunday, September 17, 2017 11:36:48 PM - Ecole Nationale Veterinaire D'Alfort IP Address:194.254.35.128

26 Haley SM, Osberg JS. Kappa coefficient calculation using multiple ratings per subject: a special communication. Phys Ther. 1989;69(11):970–4. Medline:2813523

27 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74. http://dx.doi.org/10.2307/2529310. Medline:843571

28 Costa Santos C, Costa Pereira A, Bernardes J. Agreement studies in obstetrics and gynaecology: inappropriateness, controversies and consequences. BJOG. 2005;112(5):667–9. http://dx.doi.org/10.1111/j.1471-0528.2004.00505.x. Medline:15842294

29 Keene BW. Towards evidence-based veterinary medicine. J Vet Intern Med. 2000;14(2):118–9. http://dx.doi.org/10.1111/j.1939-1676.2000.tb02223.x. Medline:10772480

30 Dawes J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. Int J Market Res. 2008;50(1):61–77.

31 Albaum G. The Likert scale revisited: an alternate version. J Market Res Soc. 1997;39:331–49.

32 Raaijmakers QAW, van Hoof A, 't Hart H, et al. Adolescents' midpoint responses on Likert-type scale items: neutral or missing values. Int J Public Opin Res. 2000;12(2):209–17. http://dx.doi.org/10.1093/ijpor/12.2.209.

33. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378–82. http://dx.doi.org/10.1037/h0031619.

34 Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. Stat Med. 1994;13(23–24):2465–76. http://dx.doi.org/10.1002/sim.4780132310. Medline:7701147

35 Burdock EI, Fleiss JL, Hardesty AS. A new view of inter-observer agreement. Person Psychol. 1963;16(4):373–84. http://dx.doi.org/10.1111/j.1744-6570.1963.tb01283.x.

36 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342(25):1887–92. http://dx.doi.org/10.1056/NEJM200006223422507. Medline:10861325

37 Scales CD Jr, Norris RD, Keitz SA, et al. A critical assessment of the quality of reporting of randomized, controlled trials in the urology literature. J Urol. 2007;177(3):1090–5, discussion 1094–5. http://dx.doi.org/10.1016/j.juro.2006.10.027. Medline:17296417

38 Maier W, Möller HJ. Metaanalyses—highest level of empirical evidence? Eur Arch Psychiatry Clin Neurosci. 2005;255(6):369–70. http://dx.doi.org/10.1007/s00406-005-0607-5. Medline:16382375

AUTHOR INFORMATION

Céline Simoneit, Clinic for Animal Reproduction, Faculty of Veterinary Medicine, Freie Universität Berlin, Königsweg 65, Haus 27, 14163 Berlin, Germany. E-mail: celine.simoneit@hotmail.de. She is interested in evidence-based veterinary medicine, small-animal reproduction, and livestock reproduction.

Wolfgang Heuwieser, Clinic for Animal Reproduction, Faculty of Veterinary Medicine, Freie Universität Berlin, Königsweg 65, Haus 27, 14163 Berlin, Germany. E-mail: heuwieser.wolfgang@vetmed.fu-berlin.de. His areas of interest are evidence-based veterinary medicine, livestock reproduction, and animal welfare.

Sebastian P. Arlt, Clinic for Animal Reproduction, Faculty of Veterinary Medicine, Freie Universität Berlin, Königsweg 65, Haus 27, 14163 Berlin, Germany. E-mail: arlt@bestandsbetreuung.de. His area of research are evidence-based veterinary medicine and small-animal reproduction.

APPENDIX 1: CHECKLIST TO ASSESS THE QUALITY OF PUBLICATIONS

Material and methodology	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not determined
1. The objective of the study is presented	□	□	□	□	□	□
2. Following information about the animals is given						
a) Number of animals	□	□	□	□	□	□
b) Inclusion criteria	□	□	□	□	□	□
c) Housing	□	□	□	□	□	□
d) Breed	□	□	□	□	□	□
e) Age	□	□	□	□	□	□
f) Sex	□	□	□	□	□	□
g) Preconditions and pretreatments	□	□	□	□	□	□
→ Description of material is clear and detailed	□	□	□	□	□	□
3. The examinations are described in detail	□	□	□	□	□	□
4. Treatments						
a) Information about the remedies or interventions (pharmaceutical agents, trade name, manufacturer) are given	□	□	□	□	□	□
b) The application of the remedy (pharmaceutical form, dose, treatment intervals) or conduction of interventions are described	□	□	□	□	□	□
→ Information about treatments or interventions are adequate	□	□	□	□	□	□
5. The monitoring is described in detail	□	□	□	□	□	□
6. Results are presented completely	□	□	□	□	□	□
7. Results are discussed critically	□	□	□	□	□	□

Study design	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not determined
1. The study was controlled	□				□	□
2. The study was randomized	□				□	□
3. The study was blinded	□				□	□
4. Study design is described in detail regarding						
a) Prospectivity/Retrospectivity	□	□	□	□	□	□
b) Adequate control group (No. of animals, comparability)	□	□	□	□	□	□
c) Type of randomisation	□	□	□	□	□	□
d) Type of blinding	□	□	□	□	□	□
→ Description of study design is clear and detailed	□	□	□	□	□	□

Statistics	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not determined
1. Statistical tests are adequate	□	□	□	□	□	□
2. Sample size is adequate	□	□	□	□	□	□
3. Level of significance is adequate	□	□	□	□	□	□
4. Handling of missing data is adequate and comprehensible	□	□	□	□	□	□
5. Analysis of data is adequate (intention-to-treat-analysis/ per-protocol-analysis, drop-out-analysis)	□	□	□	□	□	□
6. Description of statistics is adequate and comprehensible	□	□	□	□	□	□

Presentation and information content	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not determined
1. The article is written objectively	□	□	□	□	□	□
2. The summary represents the content sufficiently	□	□	□	□	□	□
3. Other studies dealing with the topic are discussed	□	□	□	□	□	□
4. The bibliography is adequate (comprehensive, current)	□	□	□	□	□	□

http://jvme.utpjournals.press/doi/pdf/10.3138/jvme.1111.113R - Sunday, September 17, 2017 11:36:48 PM - Ecole Nationale Veterinaire D'Alfort IP Address:194.254.35.128

Practical applicability	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Not determined
1. Information is relevant for practice or science	□	□	□	□	□	□
2. Applicability is discussed (techniques, equipment and knowledge, costs)	□	□	□	□	□	□
3. Alternatives are discussed	□	□	□	□	□	□
4. Side effects, limitations and complications are discussed	□	□	□	□	□	□

Conclusions

1. Data is sufficient to draw valid conclusions

   □ strongly agree □ agre □ neutral □ disagree □ strongly disagree

2. Level of evidence of the article

   □ Meta-analysis □ Randomized controlled trial □ Controlled trial without randomization

   □ Not experimental descriptive trials □ Case report □ Personal experience, expert opinion

http://jvme.utpjournals.press/doi/pdf/10.3138/jvme.1111.113R - Sunday, September 17, 2017 11:36:48 PM - Ecole Nationale Veterinaire D'Alfort IP Address:194.254.35.128