

Review

Clinical Scores in Veterinary Medicine: What Are the Pitfalls of Score Construction, Reliability, and Validation? A General Methodological Approach Applied in Cattle

Sébastien Buczinski ^{1,2,*}, Antonio Boccardo ³ and Davide Pravettoni ³

¹ Département des Sciences Cliniques, Faculté de Médecine Vétérinaire, Université de Montréal, Saint-Hyacinthe, QC J2S 2M2, Canada

² Centre d'Expertise et de Recherche Clinique en Santé et Bien-Etre Animal (CERCL), Faculté de Médecine Vétérinaire, Université de Montréal, Saint-Hyacinthe, QC J2S 2M2, Canada

³ Dipartimento di Medicina Veterinaria, Università degli Studi di Milano, via dell'Università 6, 26900 Lodi, Italy; antonio.boccardo@unimi.it (A.B.); davide.pravettoni@unimi.it (D.P.)

* Correspondence: s.buczinski@umontreal.ca; Tel.: +1-450-773-8521 (ext. 8675)

Simple Summary: Clinical scores are practical tools that can be used in the daily management of cattle. Score building and validation are a challenge involving various methodological and statistical issues. This article provides a specific framework for clinical score building where the target condition can be assessed directly or indirectly. Practical examples are given throughout the manuscript in order to build new scores or to assess score robustness.

Citation: Buczinski, S.; Boccardo, A.; Pravettoni, D. Clinical Scores in Veterinary Medicine: What Are the Pitfalls of Score Construction, Reliability, and Validation? A General Methodological Approach Applied in Cattle. *Animals* **2021**, *11*, 3244. <https://doi.org/10.3390/ani11113244>

Academic Editor: Irina Garcia-Isperto

Received: 27 September 2021
Accepted: 9 November 2021
Published: 13 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Abstract: Clinical scores are commonly used for cattle. They generally contain a mix of categorical and numerical variables that need to be assessed by scorers, such as farmers, animal caretakers, scientists, and veterinarians. This article examines the key concepts that need to be accounted for when developing the test for optimal outcomes. First, the target condition or construct that the scale is supposed to measure should be defined, and if possible, an adequate proxy used for classification should be determined. Then, items (e.g., clinical signs) of interest that are either caused by the target condition (reflective items) or that caused the target condition (formative items) are listed, and reliable items (inter and intra-rater reliability) are kept for the next step. A model is then developed to determine the relative weight of the items associated with the target condition. A scale is then built after validating the model and determining the optimal threshold in terms of sensitivity (ability to detect the target condition) and specificity (ability to detect the absence of the target condition). Its robustness to various scenarios of the target condition prevalence and the impact of the relative cost of false negatives to false positives can also be assessed to tailor the scale used based on specific application conditions.

Keywords: cattle; clinical scores; methodological approach

1. Introduction

Clinical scoring is used for various purposes, including specific conditions such as diarrhea, respiratory disease, lameness, and body condition score), and for various target applications e.g., pre-weaned calves, post-weaned calves, adult cows. A scoring system enables the scorers, including farmers, technicians, and veterinarians, to adopt a systematic approach.

With the progress of on-farm technology and precision medicine, the benefits of these human-based clinical scoring systems have been debated [1–3]. These clinical scoring systems are still used to validate several machine learning and precision medicine algorithms [3–5]. It is therefore important to understand their strengths and limitations when using

them as reference standard tests or when relying on their use to determine specific morbidity events on farms. Although the process of clinical scoring system building can be similar to the building of scales used in various psychometric fields, there are some important differences in the way they are constructed and validated, as recently reviewed [6]. For this reason, the approach proposed in this paper is tailored to clinical scoring systems that can be used for food animals.

There are several scoring systems for relatively similar purposes. For example, at least three different scoring systems are routinely used to evaluate respiratory diseases in pre-weaned dairy calves [7–9] or in veal calves [10], as indicated in Figure 1.

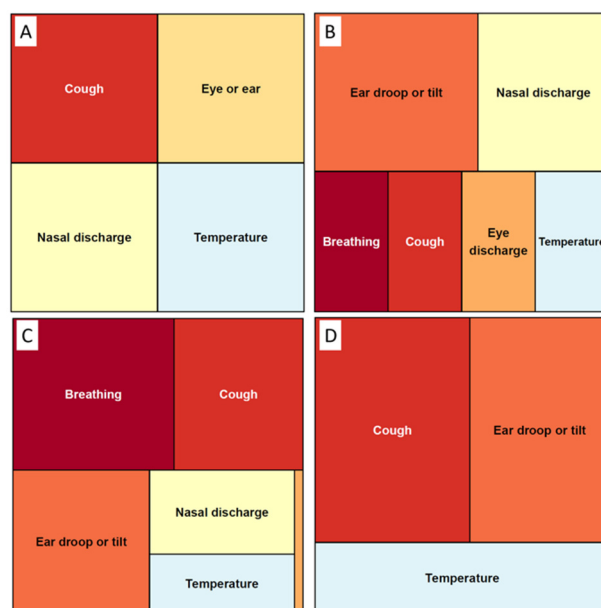


Figure 1. Treemap of the relative weights of different clinical signs included in various scoring systems used to detect bovine respiratory disease complex in calves. The size of the square is proportional to the specific clinical sign weight in the Wisconsin calf respiratory score (A) [7], California score (B) [8], Québec modified California score (C) [9], and Québec veal calf respiratory score (D) [10]. The map clearly shows the variable signs included and the relative impact on the total score of the same clinical sign.

Other scoring systems exist, for example, to determine calf diarrhea [7], lameness [11], body condition [12], and pain [13]. In these cases, it is important to know the pros and cons of the various scoring systems in order to use the best one for a specific setting. On the other hand, when there is no specific scoring method, it is important to know how to construct a scoring application. This manuscript highlights the possible objectives of a scoring system and the methodological procedure that should be constructed and validated, including the adequation between the score and what it is supposed to measure together with the reliability issues (Figure 2).

Section 2 defines the notion of construct, which is the condition or status that the score is supposed to measure. In Section 3, we highlight the problems of measuring the different items that make up the score and look at how to determine the optimal combination used in the score. Section 4 outlines the determination of score accuracy and level of discrimination. Finally, Section 5 focuses on the most difficult constructs to define which involve similar challenges as those found in psychometric and behavioral fields.

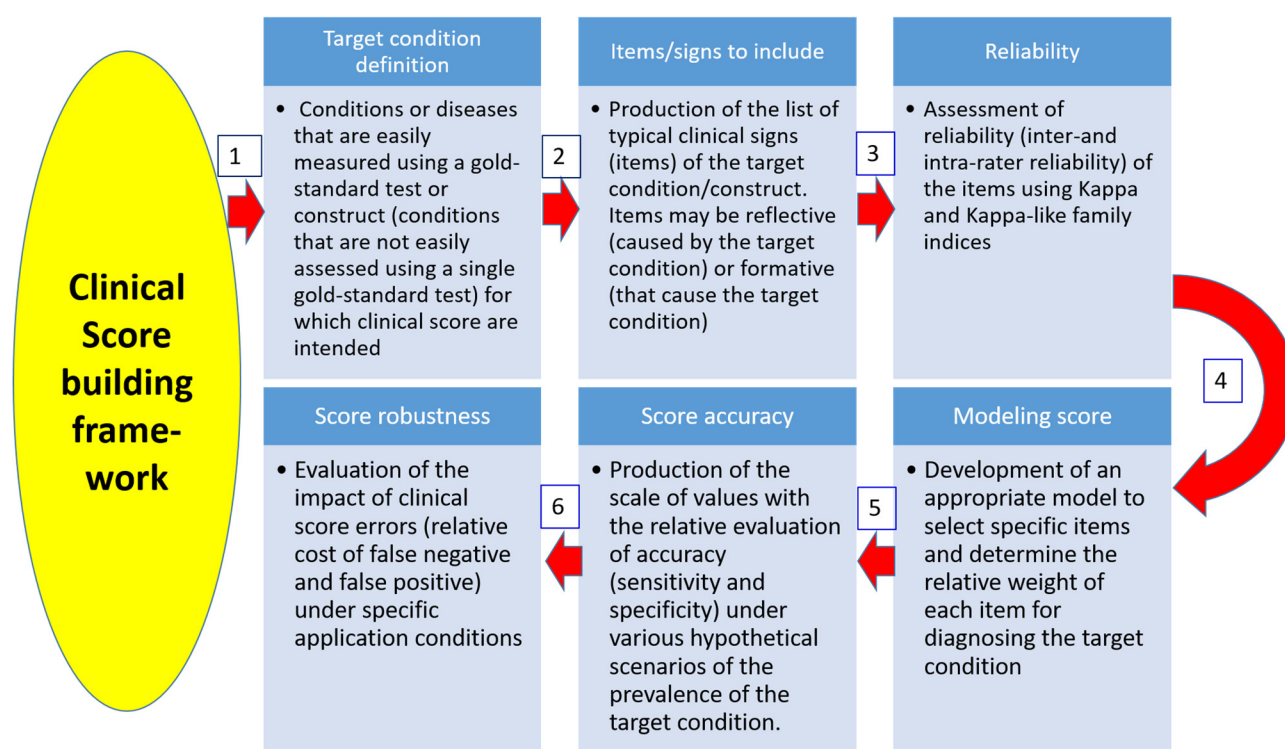


Figure 2. General framework to establish and validate a clinical scoring system in veterinary medicine. Our 6-step framework is depicted with an initial definition of the target condition (1), the potential items of interest (2), and their reliability (3). The modeling approach then determines the adequate repartition between and within item weights (4). The accuracy of the score is determined in step (5). The optimal threshold selection is then adapted to the specific conditions where the test is intended to be used (6). Score users also need to be adequately trained.

2. Condition That Needs to Be Measured: Notion of Construct

Scoring systems are generally designed with an initial target condition to diagnose that can be measured objectively using a gold-standard test. This gold standard test cannot generally be used routinely due to cost, invasiveness, convenience, or the need for a specialized lab. This is why a more practical first decision-making step is needed. For example, the lameness score based on visual gait analysis could be initially compared with the weight repartition forces on the ground and specific gait analysis measurement [14]. Again, the neonatal calf diarrhea scoring scale can be objectively determined in calves using metabolic cages that assess fecal output and dry fecal matter [15]. Other target conditions that need to be measured may be more difficult to describe accurately. This is also common in human medicine, e.g., psychology. In this situation, the different scores are used to assess a “construct” [16].

The term construct is generally used for non-easily observable characteristics in order to identify the subject of measurement. Various types of constructs where no specific affordable gold standard reference exists are also encountered in cattle. The bovine respiratory disease complex (BRD), which is a complex mixture of infectious agents affecting the upper and lower respiratory tract, is a common example in cattle medicine [17]. Stress conditions and the welfare status of animals are also construct categories that are complex to determine and define precisely [3].

The first challenge is thus when the construct cannot be easily measured or defined. In this case, the construct is a “latent variable” that cannot be measured accurately. Each scale or score aims to regroup different items that can be determined objectively or subjectively and are associated with the construct. The construct validity is challenging to define. As a rule, construct validity determines the degree to which a specific test successfully measures what it claims to test. It can, for example, be based on a specific panel determination that assesses how adequately the items included in the scale are associated

with the construct to be measured. In terms of diagnosing bronchopneumonia in calves, an expert panel can determine several clinical signs that may be of interest for the clinical diagnosis.

It is then crucial to determine other components for the validity of the score before its practical implementation. Validity is difficult to assess since it encompasses various dimensions. The content validity of the score is different from the construct validity. The content validity assesses whether the test is fully representative of what it is supposed to measure. The difference between construct and content validity is not easy to distinguish. However, the construct validity is associated with the definition of the construct that the test is supposed to measure. On the other hand, the content validity examines how the test dimensions fully represent the construct characteristics [16].

When the score aims to evaluate a more complex construct that is difficult to determine with a reference standard test (e.g., pain or stress), it is impossible to determine its accuracy using the method previously presented. A practical example of this is a score that assesses the stress or welfare of calves. In these situations, the construct has multidimensional aspects in the sense that various environmental conditions and animal-based aspects (clinical signs, behaviors, or biomarkers) may be indicative of stress without a specific combination of tests considered as a gold standard test.

Another dimension not accounted for in the literature is the quantitative assessment of the level of stress or welfare. In this case, the condition of interest should be gradated, which is another specific aspect in this condition (vs. tests used for detecting a disease or a condition (present vs. absent)). The steps for building this type of scale are similar (with some adaptations) to those in the social and behavioral fields [6]. The three steps for building the scale are item development and then scale development and validation. Item development aims to identify the domains that need to be assessed for the construct assessment. Then, specific items are built that can be based on biological and behavioral assessments. Once these items have been developed, content validity is helpful to determine the adequation between the item and the domain of interest. The content validity examines the content representativeness and relevance, as well as the technical quality of measurement. These aspects should be assessed by experts (ideally different ones from the initial team building the scale) using various quantitative and qualitative tools. The scale is then developed after pretesting questions and administering the scale in a representative sample of target animals.

3. How to Measure Items Present in the Score?

First, as with any measurement, the following should be defined: (i) the object of measurement (in this specific case the calf or the cow), (ii) the property/item that is measured (weight, rectal temperature, body condition score, specific behavior), and (iii) the values that could be used to determine this property/item (e.g., kg, °C, different strata of body condition score (BCS)). Objective items can be measured without a clinician's or operator's interpretation. For example, calf weight determined with a scale is an objective measurement that only depends on the scale's reliability. Nevertheless, the same variable (weight) can be associated with other sources of measurement error if, for example, weight is estimated using heart girth measurement [18]. In this case, in addition to tape measurement errors, other sources of uncertainty could be related to the relationship between the heart girth and body weight and how the operator applies the tape. Some items are obviously subjective (i.e., they depend on the rater), such as evaluating the body condition score in cows even using guidelines. However, many objective items are, in fact, partly subjective. For example, the rectal temperature measurement may not only be based on the type of thermometer used, but also on how the probe is inserted [19], which raises the same type of concern as tape application for heart girth measurement. Imaging methods are also affected by operator technique and image interpretation, such as quantifying lung lesions using thoracic ultrasonography in cattle [20]. This is another example of apparently objective measurements (i.e., quantifying lung lesions with a specific measure) that

are in reality subjective measurements when the operator is potentially an important source of error. However, this does not mean that subjective measurements should be excluded from scale construction, but that their intra- and inter-rater reliability should be accounted for when developing the score and scale.

The conceptual framework that illustrates how the items are related to the construct to measure is critically important. The two different frameworks used are either reflective or formative (Figure 3). In a reflective model, the items are a specific manifestation of the construct to be measured. With a formative model, the construct is the result of the item to be evaluated. A practical way of determining whether an item is formative or reflective is to determine whether a change in the construct would be associated with a change in the item manifestation. Complex constructs can be based on a mixed formative and reflective model. As a practical example in a scale assessing BRD in weaned calves, Maier et al. [21] included reflective items (sunken eyes, cough, breathing pattern, and rectal temperature) and formative items (e.g., body condition scoring, diurnal temperature fluctuation). A change in BRD status or severity might easily impact on reflective items, whereas formative items would not (Figure 3). This difference in items is not clear-cut since some formative items may in fact be reflective (in the previous case, the body condition score can be viewed as either a formative item (with low BCS calves at higher risk of being affected) or reflective (duration and severity of the respiratory condition lead to a lower BCS)).

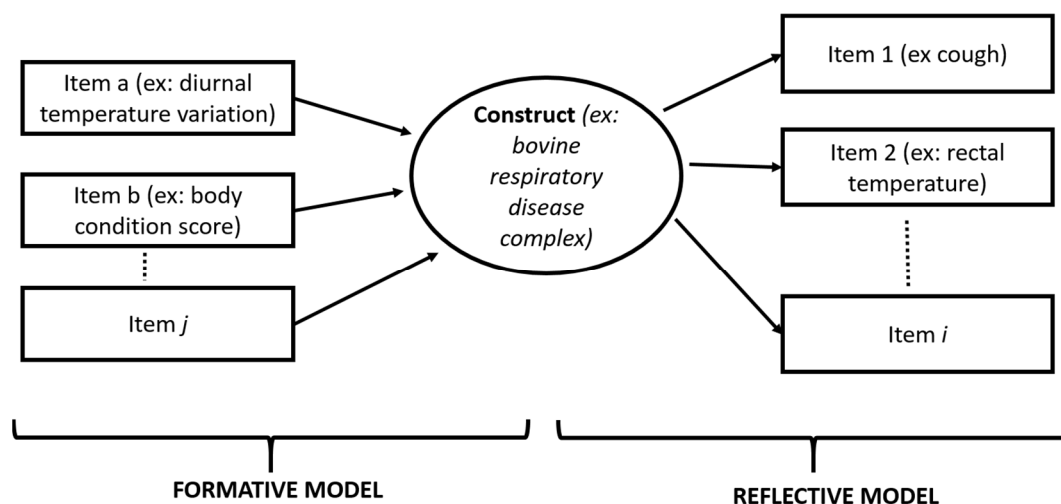


Figure 3. General framework to determine the specific items to be included for a specific construct or disease. Different items can be selected to assess a specific construct or determine a target condition. Reflective items are generally a reflection of the target condition or part of the description of the condition. In this example, cough or rectal temperature could be considered as part of the description of signs of respiratory disease (or the construct can cause them). By contrast, formative items can be considered as more causal items. They can cause the construct. In this example, diurnal variation of external temperature could be associated with an increased risk of bovine respiratory disease complex (BRD) but is not actually caused by BRD. For some items, it may be more difficult to determine both their formative and reflective roles. For example, the BCS could either be formative (e.g., calves with a low BCS could be at higher risk of being sick) or reflective (long disease duration may lead to a low BCS).

3.1. Items Selected and Classification of the Different Possible Categories of Items

The association between the different clinical signs or factors included in the score needs to be detailed. For example, adding highly correlated clinical signs would not be useful from a monitoring basis. If one clinical sign is highly correlated with another, knowing one of the clinical signs would implicitly mean that the second sign is also known (because of the high correlation between the signs). The information gained from knowing one rather than two signs would therefore not improve the monitoring capacity in a large

population. However, using correlated questions could be essential for complex constructs with various dimensions in order to gather information that differentiates between some patients and highlights a single or multidimensional domain of the score [6].

A precise selection of the most practical signs is essential as targeted score-users measure their reliability. The stratification of the items to be scored is another major topic of interest. The stratification process should be clearly defined with clear mutually exclusive definitions. Several items scored in various cattle scoring systems do not meet these criteria. In the Wisconsin score system, nasal discharge is noted as normal serous discharge (score = 0), small amount of unilateral cloudy discharge (score = 1), bilateral, cloudy, or excessive mucus discharge (score = 2), and a copious bilateral mucopurulent discharge (score = 3). What would happen if a calf only showed a small amount of mucous discharge? And what is a “small amount” of unilateral cloudy discharge, or how should “copious” be objectively defined? Some raters will not have the same conception of “small amount” or “copious.” Pictures associated with the scorecards that explain the score are helpful but not self-explanatory enough to ensure that the definition is reliable within and between raters. This may explain a significant part of the score variability that is not attributable to the measurement object (calf) but to the limited reliability between scorers.

This is only one example of the reliability issues when implementing scoring systems in practice, especially when the score is a specific outcome for monitoring different interventions or management changes. Improvement in definitions using more objective measurements (e.g., the quantification of the importance of the discharge) or the duration of the observation period to define cough frequency could be used to decrease rater subjectivity. Specific training can also improve consistency among scorers.

Using multiple categories for one item assessment is also useful for the gradation of clinical sign modifications. The natural drawback of using multiple categories is that it is associated with an inflated risk of altered reliability due to more choices in the item scale [11]. This should also be accounted for when proposing the different categories of items. Another challenge is to determine how the clinical sign is associated with the construct, its relative importance in relation to other signs included in the score, and the relative weight of the different categories of the scale within the same clinical sign.

3.1.1. Intra and Inter-Rater Reliability

Reliability is a key issue for any clinical measurement. When the various items to be rated have been selected, their test-retest reliability (first by the same rater and then by different raters) should be assessed. Depending on the type of item to be rated, several complementary approaches can be used [22]. Kappa and kappa-like family indices are generally used to differentiate agreements that occur purely by random chance versus true agreements. The difference between the multiple types of agreement indices is generally based on different definitions in agreement by chance as well as the way a partial agreement (in ordinal scale) is accounted for [23]. We only focus on the most common parameters used.

3.1.2. Two Raters Categorical Scale

Cohen’s kappa [24] is by far the most commonly reported agreement parameter in veterinary medicine and many other medical fields. The general framework for kappa calculation is to correct the percentage of agreement between the two raters corrected with the percentage of the agreement only due to chance (i.e., the percentage of agreement that would be obtained if the raters’ ratings were independent of each other). This concept is explained in Table 1. When n total subjects are scored by two raters (1 and 2), the results of the tests can either be (1,0), the subjects are cross-classified based on the test results probability with r_{ab} (a = test result by rater 1 (0 or 1), b = test result by rater 2 (0 or 1)) and $r_{11} + r_{10} + r_{01} + r_{00} = 1$.

Table 1. Example of 2×2 table results based on a dichotomous scale obtained by two raters scoring n different animals for an item.

	Test = 1 for Rater or Test 2	Test = 0 for Rater or Test 2	
Test = 1 for rater or test 1	$n_{11} = n \times r_{11}$	$n_{10} = n \times r_{10}$	
Test = 0 for rater or test 1	$n_{01} = n \times r_{01}$	$n_{00} = n \times r_{00}$	n

The observed raw probability of agreement (P_a) between the two raters is

$$P_a = (r_{11} + r_{00})$$

The chance agreement (P_c) is defined as the probability of observing agreement if the raters were independent:

$$P_c = (r_{11} + r_{10}) \times (r_{11} + r_{01}) + (r_{01} + r_{00}) \times (r_{10} + r_{00})$$

The kappa is then simply calculated using the following formula correcting P_a with P_c .

$$K = \frac{P_a - P_c}{1 - P_c}$$

The standard deviations for calculating 95% CI using the frequentist approach is [24]:

$$SD(K) = \sqrt{\frac{P_a(1 - P_a)}{(1 - P_c)^2}}$$

The 95%CI of the K value can therefore be obtained using the standard normal distribution of the K with $K \pm 1.96$ SE when α error is 0.05. This can further be extended for other values of α with the $100 \times (1 - \alpha)$ confidence intervals for K computed as:

$$K \pm z_{\alpha/2} \times \frac{SD(K)}{\sqrt{n}}$$

The interpretation of the confidence intervals is not clinically intuitive and is generally falsely interpreted as Bayesian credible intervals. Determining K confidence intervals has been criticized from a frequentist perspective [25]. Benchmarking these intervals is not evident under the frequentist framework despite some recently proposed approaches [23,26].

These concepts can be easily illustrated when two raters scored a population of 100 calves using a specific scoring system (see supplementary file). It is straightforward to extend the kappa calculation for multiple categories (>2). In this situation, a weighted kappa (K_w) can also be calculated if the categories are ordinal, indicating that a different weight is applied depending on putting more weight on the error in two distant categories versus differences in adjacent categories [23].

The Bayesian version of the K with readily applicable credible intervals calculated without specific parametric assumption can also be obtained using the same equation for calculating K but considering that each cell content of the 2×2 table has a gamma (Γ) distribution $\Gamma(k, \theta)$ with shape k and relatively large scale θ . The shape is directly obtained from the numbers in the initial classification ($n_{11}, n_{10}, n_{01}, n_{00}$), adding 1 ($n_{11} + 1, n_{10} + 1, n_{01} + 1, n_{00} + 1$) since these variables are assumed to follow a Dirichlet distribution. The scale θ is generally given a >1 value (e.g., 2), which helps to extend the ranges of possible distribution. The K values are then obtained from multiple iterations of the process to obtain the density of K distribution [27].

The Bayesian version of agreement and reliability measures is currently quite complex as there is no easy-to-use calculator (supplementary file). The reliability measures

generally rely on more extensive coding ability and, although promising, are only used by few researchers [27,28].

3.1.3. Multiple Raters

The agreement process can be extended with multiple-rater scoring by generalizing the previous concept of agreement by chance based on Fleiss's extension of Scott's π statistics (known as Fleiss' kappa K_F) [23]. Scott's π statistic is slightly different from K since it accounts for the mean predicted probability between r different raters ($r \geq 2$) scoring an object in q different categories.

$$K_F = \frac{Pa - Pc|\pi}{1 - Pc|\pi}$$

With:

$$Pa = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^q \frac{r_{ik} (r_{ik} - 1)}{r(r-1)}$$

where r_{ik} is the proportion of the r raters affecting to the i^{th} subjects the k categories and with:

$$Pc|\pi = \sum_{k=1}^q \hat{\pi}_k^2$$

and

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^n \frac{r_{ik}}{r}$$

The confidence intervals can be obtained from a specific determination K_F standard error (for details see [23]). Although kappa-like indices are still widely used, they have various drawbacks known as kappa paradoxes, especially when the calculated Pc is very high and where K is close to 0 despite a high crude percentage of agreement [29]. New indices, such as Gwet's agreement coefficients (AC), have been explored to avoid the pitfalls involved in calculating Pc [23]. The AC parameters complement traditional K statistics [26,30].

There are many other agreement/reliability indices, but they are beyond the scope of this article. However, a common situation when testing the reliability of categorical item measurement is missing data, e.g., the different scoring sessions where different raters assess different animals [31]. In this specific case, not all the rater-animal pairs are available. Since they only use the complete dataset, calculations of kappa and kappa-like statistics (e.g., Fleiss kappa) are not robust to missing data. In this situation, Krippendorff's alpha is the most robust approach given that it does not only exploit the complete pair-rater scores [32]. See [23,32] for information on how to calculate this reliability index and associated confidence intervals. The Gwet AC parameters are also robust to missing data.

In summary, the agreement assessment between raters assessing a categorical scale is complex and can generally not be performed simply by reporting one parameter. Reporting multiple aspects of the agreement with raw agreement and chance-corrected agreement parameters is recommended. However, the specific association of which chance-adjusted agreement parameter to report is still an open issue.

3.1.4. Numeric Items

The previous calculations formerly adapted to categorical scales do not work for quantitative items that can be measured to assess the health of calves or cows. In these situations, the measured item (M) is generally a numerical value that quantifies the true

value of interest (η) with a specific error (ε). This relationship can therefore be written as follows:

$$M = \eta + \varepsilon$$

The specific ε term can further be classified as the error or measurement device, error due to operator, or any other cause of error. Reliability concepts applied to these measurements are associated with the variance (σ^2) of i^{th} repeated measurements M by different raters, time or specific conditions (M_i), and the specific error term (ε_i). For example, the true internal temperature (η) can be assessed by the rectal temperature measurement (M_i) by i different persons or using i different thermometers. The aim of the reliability parameters is generally to quantify the variance of the true measure ($\sigma^2(\eta)$) versus the general variance of the error ($\sigma^2(\varepsilon_i)$), which can further be simplified as:

$$\text{Reliability} = \frac{\sigma^2(\eta)}{\sigma^2(\eta) + \sigma^2(\varepsilon_i)} = \frac{\sigma^2_{\text{patient}}}{\sigma^2_{\text{patient}} + \sigma^2_{\text{error}}}$$

Therefore, the reliability assesses the relative error versus the true value variance, or the ratio of the variance of interest divided by the variance of interest plus the unwanted (noisy) variance. This family of reliability parameters is also called intra-class correlation (ICC) measures [33]. There are multiple types of ICC that depend on how error variance is partitioned. These include combinations between one-way or two-way random, two-way mixed effects, and absolute (ICC(A)) vs. consistency (ICC(C)) agreement (for details see [33]).

3.1.5. Benchmarking Reliability Parameters

Once a reliability value has been obtained for either categorical or numerical items, it is crucial to determine whether these values are compatible with the intended use of the score or item assessment. Most benchmarks have not been extensively validated but are opinion-based [34]. For example, is not easy to know whether in practice a kappa of 0.8 is different from a kappa of 0.6. Based on the calculation difference, using a single agreement reliability parameter is not enough to assess all the dimensions of agreement. Several differences may be observed and should be discussed regarding the intended use of the score or item [23]. Although general benchmarks have been reported for numerical items using ICC measures [33], they have not been thoroughly validated and should be used in the light of the specific study context.

4. How Can the Best Combination of Items Be Determined for Use in a Score?

4.1. Score Building to Determine the Relative Weight between and within Items That Are Assessed to Measure a Simple Construct

Once different clinical items included in the score have been validated with their inter-and intra-rater variability, the next step is to determine the strength of association between all items and the different categories with the construct to be measured by the different items and their values. Linear, generalized linear model, or survival analyses possibly accounting for correlations between the patients (e.g., herd aggregation level) should thus be built. A specific framework for diagnostic or prediction models has been proposed with the transparent reporting of a multivariable prediction model for individual diagnosis or prognosis (TRIPOD) statements [35]. The general framework of reporting can be explained for a specific outcome Y , with $X_1, \dots, X_n$ clinical signs or specific covariates with respectively ($k, \dots, n_k$ categories) and u the random error (if any) in the form:

$$g(Y) = \beta_0 + \beta_{11}X_{11} + \dots + \beta_{1k}X_{1(k-1)} + \dots + \beta_{n1n}X_{n1} + \dots + \beta_{nk_n}X_{n(k_n-1)} + u$$

The link function between the construct or condition to diagnose (Y) and the different covariates depends on the type of construct and the way it is assessed. When the construct

is a binary condition that can be diagnosed with an accurate reference standard test (considered as a gold standard), the binomial link with logistic regression to determine the probability of the presence of the disease (p , which is the probability that the gold standard test is positive) is the natural choice, and this was exploited in a study using a specific BRD case definition [8]. Survival analysis models are generally recommended when the time to event outcome is of interest.

In many clinical conditions, there is no affordable 100% accurate diagnostic test to determine the animal status (Y). For example, in the diagnosis of BRD where, except for necropsy, no reference standard test is accurate enough to be exempt from classification error. In these conditions, there are three alternatives: (i) choosing a proxy of the target condition which is at risk of bias due to imperfect accuracy (e.g., for BRD: thoracic ultrasonography); (ii) using a composite reference standard test can be a solution (e.g., for BRD, using an association between thoracic ultrasound, serum haptoglobin, and bronchoalveolar lavage) although this has its own risk of bias [36]; and finally (iii) accounting for the imperfect accuracy of a test that can be done in a practical setting to define the latent status of the animal [37]. The modeling approach is similar in the first two conditions since the reference test is defined as positive ($Y = 1$) and negative ($Y = 0$) animals.

However, the framework is different when the true status (target condition or construct) is considered unknown (latent) but can be described based on the known accuracy of the imperfect test used. In this case, the probability that an animal has an outcome ($p = p(Y = 1)$) can be associated with the probability that the animal reference standard test is positive (pT+) or negative, accounting for test sensitivity (Se) and specificity (Sp) as follows:

$$g(Y) = \text{logit } p = \beta_0 + \beta_{11}X_{11} + \dots + \beta_{nk_n}X_{nk_n}$$

$$p_{T+} = p * Se + (1 - p) * (1 - Sp)$$

Therefore, the imperfect accuracy of the test means that unbiased model characteristics can be determined when the outcome is measured with error [8,38].

Internal vs. External Validation of the Model

The model needs to be validated especially in terms of refining the estimates of β parameters in order to account for potential shrinkage when applied to a new population. This topic is also a fundamental research area in biostatistics that encompasses the current manuscript [39,40]. Initial model selection is always associated with overestimating model accuracy and discrimination. This is because the model performance is higher in the dataset where the original model was developed than in a new dataset (“testimation” problem).

The value of the different β should then be the basis for calculating the individual weights of model predictors. It is essential not to use the transformation of the estimates as a score weight (e.g., OR) because linearity between the predictors and the construct is only valid on the assumption of a $g(.)$ link and not with the odds ratio derived from the coefficients in a logistic regression approach [41]. For practical purposes, the β are generally multiplied by a common number (e.g., 5 or 10) and rounded to obtain the full score and items.

4.2. Measurement of Complex Construct

Measuring a complex construct with a multidimensional aspect is totally different from the precedent approach. For example, building a tool to assess the welfare of calves is not comparable with a standard reference test, which would only capture one specific dimension of the construct. Measuring serum cortisol, for example, as an indicator of stress (e.g., proxy of the target condition to measure) would be a welfare dimension. However, the definition of welfare includes measuring a specific biomarker. In this situation,

building the score would have a different assessment framework and other specific statistical parameters that are outside the scope of this article [6].

5. Accuracy of the Score

5.1. Case 1: The True Status of the Construct Can Be Determined (Directly or Indirectly)

Once the ideal repartition between the various items and the various categories within the same items have been determined, it is then necessary to determine how to implement the score in practice with a specific recommendation concerning the thresholds. For the diagnosis of a dichotomous condition, the accuracy of the scoring system could be determined using the ability to diagnose animals with the target condition (sensitivity) and the ability to diagnose non-affected animals (specificity).

The optimal combination of sensitivity and specificity can be determined using different decision thresholds. Minimizing the total misclassification rate, and thus maximizing the Se+Sp sum (Figure 4), is generally the default in many situations for diagnostic test accuracy assessments [42]. However, this selection is arbitrary and does not account for the test's intended use and the differential value of false-positive versus false-negative cases. In some situations, missing truly affected animals is more deleterious than the false-positive classification of non-affected animals.

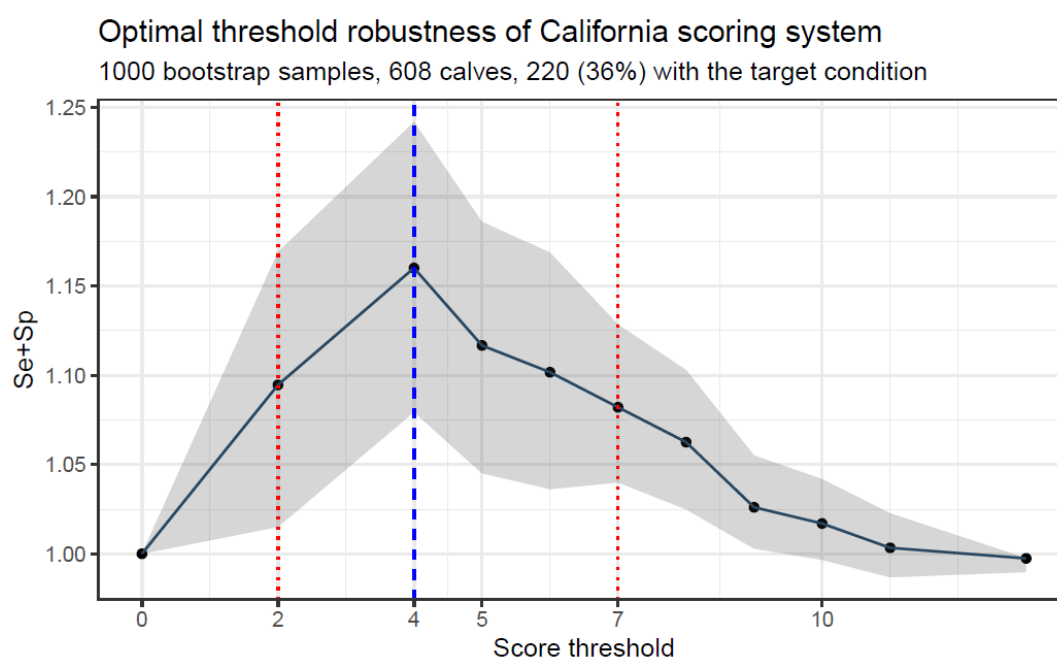


Figure 4. Determination of the optimal threshold based on a specific score. This figure represents the California respiratory score threshold for detecting ultrasonographic lung consolidation (target condition defined as a maximal consolidation depth ≥ 1 cm) in 608 pre-weaned calves from 39 different dairy herds [9]. The accuracy of the score is presented in the y-axis with the sum of sensitivity (Se) and specificity (Sp) of the score depending on the threshold ($\geq$ threshold to define a positive score, $<$ for a negative score). The dots represent the observed value of the score, and the 95% confidence interval band was obtained after 1000 bootstrap estimates. Even though the threshold of ≥ 4 is numerically associated with the highest accuracy (dashed blue line), there is no evidence that this threshold is better than any threshold between 2 and 7 (dotted red lines) due to the confidence interval width.

A specific example of this is bronchopneumonia in calves, where not finding affected animals would have a more deleterious animal health impact in terms of the risk of incomplete cure and negative outcome as against recommending treatment in a non-affected calf. This relative cost would depend on the type of cost assessed. Using the same BRD example but from a public health/rational use of antimicrobials, the associated risk of antimicrobial use in non-affected animals would have a significant impact. The exact

definition of “cost” may depend on the context (farm, country, etc.). A complete assessment of costs can therefore not be easily performed. It is thus important to test various plausible scenarios that cover broad but plausible situations (Figure 5).

Various test applications in terms of the possible prevalence of the target condition (p) and the relative cost of false-negative versus false-positive (FN:FP) cases (r) can be determined by calculating the misclassification cost term (MCT), which should be minimized [43]. The MCT calculation can be written as:

$$MCT = (1 - p) * (1 - Sp_t) + r * p * (1 - Se_t)$$

for the various possible t thresholds of decision. The MCT is not in itself an extensive cost analysis examination but rather a practical way to address the robustness of the cut-off depending on different plausible (“what if”) scenarios of the target condition prevalence and r values. If the minimum MCT area is very variable depending on p and r , the recommendation for applying a common threshold should be determined more in depth before applying the test in practice (Figure 5). When the MCT reaches its minimum independently of p and r , that threshold is considered a relatively robust threshold which can be applied in the various conditions found in practice. However, robustness is not equivalent to accuracy since a robust threshold can have a limited accuracy.

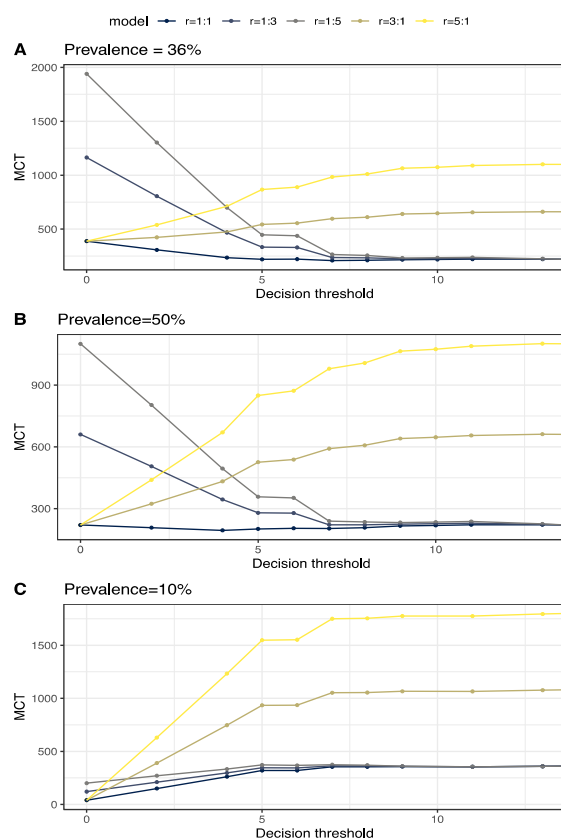


Figure 5. Misclassification cost-term (MCT) analysis of a specific score to determine lung consolidation. The same dataset as in Figure 3 was used to determine the misclassification cost term using: (A) the original prevalence of the target condition in the dataset (36% of calves were detected with ultrasonographic lung consolidation depth ≥ 1 cm); (B) a prevalence of 50%; and (C) a prevalence of 10% based on internal resampling. Different false negative (FN): false positive (FP) cost ratios (r) are indicated. In the first scenario, r was considered as 1:1, meaning that FN and FP had the same impact. Two scenarios where FN calves were considered more costly than FP calves with 3:1 and 5:1 r values. Finally, two scenarios considered that false-positive calves were more costly than FN cases with 1:3 and 1:5 r values. The 5 MCT curves are then derived to determine the score threshold robustness for various settings. The optimal threshold minimizes the y-axis value.

5.2. Case 2: More Complex Construct

Because it is beyond the scope of this paper, we have not discussed the best way to deal with the items depending on the nature of their quantification, the dimension they are supposed to measure, and the impact of the nature of the item (i.e., formative vs. reflective) on scale building (for details see [6]).

In conclusion, clinical scores have an important role in determining various health and welfare conditions for food animals. The construction, reliability and, validation of the score depend on the construct or target condition of interest. We have focused on constructing or targeting conditions that can be determined directly (using a gold standard test) or indirectly (using an imperfect reference standard test). The reliability of the possible items to be included, the acquirement of the relative item weight in the final score, and the determination of the decision threshold to apply based on various contexts have been explained. Future work should focus on the internal and external validation of the decision threshold in a different population from where it was initially derived to improve its robustness. Although this paper has focused on bovine health and welfare condition monitoring, the proposed framework could be extended to any species of veterinary interest since the same concepts and rules would be applicable. It is also important to mention that automatic data collection is also a promising area of research in the detection of animal health problems with the implementation of a machine learning strategy and artificial intelligence technique. We chose to focus on clinical signs that can be detected by farmers or veterinarians, but automatic behavior or anomaly detection is definitely an important potential evolution of disease detection.

Supplementary Materials: The following are available online at www.mdpi.com/article/10.3390/ani11113244/s1, Supplementary file: Example of agreement calculation with frequentist method and Bayesian method.

Author Contributions: Conceptualization and writing—original draft, S.B.; writing—review and editing, A.B. and D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Carslake, C.; Vázquez-Diosdado, J.A.; Kaler, J. Machine learning algorithms to classify and quantify multiple behaviours in dairy calves using a sensor: Moving beyond classification in precision livestock. *Sensors* **2021**, *21*, 88.
2. Duthie, C.A.; Bowen, J.M.; Bell, D.J.; Miller, G.A.; Mason, C.; Haskell, M.J. Feeding behaviour and activity as early indicators of disease in pre-weaned dairy calves. *Animal* **2021**, *15*, 100150.
3. Stachowicz, J.; Umstätter, C. Do we automatically detect health-or general welfare-related issues? A framework. *Proc. R. Soc. B* **2021**, *288*, 20210190.
4. Knauer, W.A.; Godden, S.M.; Dietrich, A.; James, R.E. The association between daily average feeding behaviors and morbidity in automatically fed group-housed preweaned dairy calves. *J. Dairy Sci.* **2017**, *100*, 5642–5652.
5. Knauer, W.A.; Godden, S.M.; Dietrich, A.; Hawkins, D.M.; James, R.E. Evaluation of applying statistical process control techniques to daily average feeding behaviors to detect disease in automatically fed group-housed preweaned dairy calves. *J. Dairy Sci.* **2018**, *101*, 8135–8145.
6. Boateng, G.O.; Neilands, T.B.; Frongillo, E.A.; Melgar-Quinonez, H.R.; Young, S.L. Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Front. Public Health* **2018**, *6*, 149.
7. McGuirk, S.M. Disease management of dairy calves and heifers. *Vet. Clin. N. Am. Food Anim. Pract.* **2008**, *24*, 139–153.
8. Love, W.J.; Lehenbauer, T.W.; Kass, P.H.; Van Eenennaam, A.L.; Aly, S.S. Development of a novel clinical scoring system for on-farm diagnosis of bovine respiratory disease in pre-weaned dairy calves. *PeerJ* **2014**, *2*, e238.
9. Buczinski, S.; Fecteau, G.; Dubuc, J.; Francoz, D. Validation of a clinical scoring system for bovine respiratory disease complex diagnosis in preweaned dairy calves using a Bayesian framework. *Prev. Vet. Med.* **2018**, *156*, 102–112.

10. Berman, G.; Francoz, D.; Dufour, S.; Buczinski, S. Un Nouvel Outil Diagnostique Pour Détecter les Bronchopneumonies Infectieuses à Traiter Chez les Veaux Lourds! Available online: <http://bovin.qc.ca/wp-content/uploads/2021/03/P44-45-46-Outil-bronchopneumonies.pdf> (accessed on 1 September 2021).
11. Schlageter-Tello, A.; Bokkers, E.A.; Groot Koerkamp, P.W.; Van Hertem, T.; Viazzi, S.; Romanini, C.E.; Halachmi, I.; Bahr, C.; Berckmans, D.; Lokhorst, K. Effect of merging levels of locomotion scores for dairy cows on intra- and interrater reliability and agreement. *J. Dairy Sci.* **2014**, *97*, 5533–5542.
12. Roche, J.R.; Friggens, N.C.; Kay, J.K.; Fisher, M.W.; Stafford, K.J.; Berry, D.P. Invited review: Body condition score and its association with dairy cow productivity, health, and welfare. *J. Dairy Sci.* **2009**, *92*, 5769–5801.
13. Tomacheuski, R.M.; Monteiro, B.P.; Evangelista, M.C.; Luna, S.P.L.; Steagall, P.V. Measurement properties of pain scoring instruments in farm animals: A systematic review protocol using the COSMIN checklist. *PLoS ONE* **2021**, *16*, e0251435.
14. Van Nuffel, A.; Zwervaeagher, I.; Pluym, L.; Van Weyenberg, S.; Thorup, V.; Pastell, M.; Sonck, B.; Saeys, W. Lameness detection in dairy cows: Part 1. How to distinguish between non-lame and lame cows based on differences in locomotion or behavior. *Animals* **2015**, *5*, 838–860.
15. Perryman, L.E.; Kapil, S.J.; Jones, M.L.; Hunt, E.L. Protection of calves against cryptosporidiosis with immune bovine colostrum induced by a *Cryptosporidium parvum* recombinant protein. *Vaccine* **1999**, *17*, 2142–2149.
16. De Vet, H.C.; Terwee, C.B.; Mokkink, L.B.; Knol, D.L. Concepts, Theories and Models, and Types of Measurements. In *Measurement in Medicine: A practical Guide*; Cambridge University Press: Cambridge, UK, 2011.
17. Buczinski, S.; Pardon, B. Bovine respiratory disease diagnosis: What progress has been made in clinical diagnosis? *Vet. Clin. N. Am. Food Anim. Pract.* **2020**, *36*, 399–423.
18. Heinrichs, A.J.; Heinrichs, B.S.; Jones, C.M.; Erickson, P.S.; Kalscheur, K.F.; Nennich, T.D.; Heins, B.J.; Cardoso, F.C. Short communication: Verifying Holstein heifer heart girth to body weight prediction equations. *J. Dairy Sci.* **2017**, *100*, 8451–8454.
19. Naylor, J.M.; Streeter, R.M.; Torgerson, P. Factors affecting rectal temperature measurement using commonly available digital thermometers. *Res. Vet. Sci.* **2012**, *92*, 121–123.
20. Buczinski, S.; Buathier, C.; Bélanger, A.M.; Michaux, H.; Tison, N.; Timsit, E. Inter-rater agreement and reliability of thoracic ultrasonographic findings in feedlot calves, with or without naturally occurring bronchopneumonia. *J. Vet. Intern. Med.* **2018**, *32*, 1787–1792.
21. Maier, G.U.; Rowe, J.D.; Lehenbauer, T.W.; Karle, B.M.; Williams, D.R.; Champagne, J.D.; Aly, S.S. Development of a clinical scoring system for bovine respiratory disease in weaned dairy calves. *J. Dairy Sci.* **2019**, *102*, 7329–7344.
22. Kottner, J.; Audigé, L.; Brorson, S.; Donner, A.; Gajewski, B.J.; Hróbjartsson, A.; Streiner, D.L. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Int. J. Nurs. Stud.* **2011**, *48*, 661–671.
23. Gwet, K.L. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters*, 4th ed.; Advanced Analytics LLC: Piedmont, CA, USA, 2014.
24. Cohen, J.A. Coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46.
25. Hilden, J. Commentary: On NRI, IDI, and “good-looking” statistics with nothing underneath. *Epidemiology* **2014**, *25*, 265–267.
26. Berman, J.; Francoz, D.; Abdallah, A.; Dufour, S.; Buczinski, S. Evaluation of inter-rater agreement of the clinical signs used to diagnose bovine respiratory disease in individually housed veal calves. *J. Dairy Sci.* **2021**, in press.
27. Broemeling, L.D. *Bayesian Methods for Measures of Agreement*, 1st ed.; Chapman and Hall/CRC: Boca Raton, FL, 2020.
28. Van Oest, R.; Girard, J.M. Weighting Schemes and Incomplete Data: A Generalized Bayesian Framework for Chance-Corrected Interrater Agreement. 2021. Available online: <https://psyarxiv.com/s5n4e/> (accessed on 2 September 2021).
29. de Vet, H.C.W.; Mokkink, L.B.; Terwee, C.B.; Hoekstra, O.S.; Knol, D.L. Clinicians are right not to like Cohen’s κ . *Br. Med. J.* **2013**, *346*, f2125.
30. Walsh, P.; Thornton, J.; Asato, J.; Walker, N.; McCoy, G.; Baal, J.; Banimahd, F. Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the emergency department. *PeerJ* **2014**, *2*, e651.
31. Pardon, B.; Buczinski, S.; Deprez, P.R. Accuracy and inter-rater reliability of lung auscultation by bovine practitioners when compared with ultrasonographic findings. *Vet. Rec.* **2019**, *185*, 109–109.
32. Zapf, A.; Castell, S.; Morawietz, L.; Karch, A. Measuring inter-rater reliability for nominal data—Which coefficients and confidence intervals are appropriate? *Med. Res. Methodol.* **2016**, *16*, 1–10.
33. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163.
34. Dohoo, I.R.; Martin, S.W.; Stryhn, H. *Veterinary Epidemiologic Research*, 2nd ed.; VER. Inc., Charlottetown, PEI: Canada, 2009.
35. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Br. J. Surg.* **2015**, *102*, 148–158.
36. Schiller, I.; van Smeden, M.; Hadgu, A.; Libman, M.; Reitsma, J.B.; Dendukuri, N. Bias due to composite reference standards in diagnostic accuracy studies. *Stat. Med.* **2016**, *35*, 1454–1470.
37. Branscum, A.J.; Gardner, I.A.; Johnson, W.O. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Prev. Vet. Med.* **2005**, *68*, 145–163.
38. McInturff, P.; Johnson, W.O.; Cowling, D.; Gardner, I.A. Modelling risk when binary outcomes are subject to error. *Stat. Med.* **2004**, *23*, 1095–1109.
39. Harrell, F.E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2019.

-
40. Steyerberg, E.W. *Clinical Prediction Models*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2019.
 41. Moons, K.G.; Harrell, F.E.; Steyerberg, E.W. Should scoring rules be based on odds ratios or regression coefficients? *J. Clin. Epidemiol.* **2002**, *55*, 1054–1055.
 42. Pepe, M.S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*, 1st ed.; Oxford University Press: Oxford, UK, 2003.
 43. Greiner, M. Two-graph receiver operating characteristic (TG-ROC): Update version supports optimisation of cut-off values that minimise overall misclassification costs. *J. Immunol. Methods* **1996**, *191*, 93–94.