Meat Science 100 (2015) 232–236

Contents lists available at ScienceDirect

Meat Science

journal homepage: www.elsevier.com/locate/meatsci

# Validation of boar taint detection by sensory quality control: Relationship between sample size and uncertainty of performance indicators

Daniel Mörlein [a,*], Rune Haubo Bojesen Christensen [b], Jan Gertheiss [c,d]

[a] Department of Animal Sciences, Meat Science Group, University of Göttingen, D-37075 Göttingen, Germany
[b] Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
[c] Department of Animal Sciences, Biometrics & Bioinformatics Group, University of Göttingen, D-37075 Göttingen, Germany
[d] Centre for Statistics, University of Göttingen, D-37075 Göttingen, Germany

ARTICLE INFO

*Article history:*
Received 10 March 2014
Received in revised form 3 October 2014
Accepted 7 October 2014
Available online 16 October 2014

*Keywords:*
Experimental design
Panel performance
Quality control
Boar taint
Androstenone
Skatole

ABSTRACT

To prevent impaired consumer acceptance due to insensitive sensory quality control, it is of primary importance to periodically validate the performance of the assessors. This communication showcases how the uncertainty of sensitivity and specificity estimates is influenced by the total number of assessed samples and the prevalence of positive (here: boar tainted) samples. Furthermore, a statistically sound approach to determining the sample size that is necessary for performance validation is provided. Results show that a small sample size is associated with large uncertainty, i.e., confidence intervals and thus compromising the point estimates for assessor sensitivity. In turn, to reliably identify sensitive assessors with sufficient test power, a large sample size is needed given a certain level of confidence. Easy-to-use tables for sample size estimations are provided.

## 1. Introduction

Similar to instrumental procedures for quality control, sensory evaluation needs performance assessment for validation and further improvement (Munoz, 2002). It is of primary concern to carefully select and train panellists that are able to reliably differentiate between products being, e.g., in or out the specification. Although this has rarely been documented for sensory methods, the use of sensitivity, specificity, and accuracy are well established performance indicators for in/out problems and are, for example, applied for evaluation of clinical tests (Lalkhen & McCluskey, 2008).

At the moment the detection of so-called boar taint in male pigs requires the use of sensory quality control due to the lack of rapid instrumental techniques (Haugen, Brunius, & Zamaratskaia, 2012). EU regulations require that meat must be declared unfit for human consumption if organoleptic anomalies, such as sexual odour, occur (Regulation EC No 854/2004, 2004). It is, therefore, imperative when raising intact boars to control so-called boar-taint (Lundström, Matthews, & Haugen, 2009) which is mainly caused by elevated levels of the testicular steroid androstenone (5α-androst-16-en-3-one) and/or skatole (4-Methyl-2,3-benzopyrrole) produced microbially in the digestive tract (Patterson, 1968; Vold, 1970). Therefore, sensory evaluation of boar carcasses is performed in-line or at-line by trained assessors. Release of key volatiles is usually facilitated by heating subcutaneous fat tissue using, e.g., a hot iron, soldering iron, pyro pen or microwave (Bekaert et al., 2013; Mathur et al., 2012; Whittington et al., 2011). For sensory quality control it is often preferred to use a rather simple scoring system as compared to descriptive sensory evaluation. As regards the case of boar taint, when the assessors are judging whether the odor is present or not, this corresponds to an IN/OUT test (Munoz, 2002) or A-not A test (Brockhoff & Christensen, 2010; Macmillan & Creelman, 2005); sometimes ordinal scales are applied where assessors score the abundance of boar taint from 0 (= no taint) to 4 or 5 (= strong boar taint) with subsequent dichotomisation of the original scores (Mathur et al., 2012).

As olfactory perception varies between individuals it is therefore crucial to select assessors according to their olfactory acuity (Meier-Dinkel, Sharifi, et al., 2013; Mörlein, Meier-Dinkel, Moritz, Sharifi, & Knorr, 2013). Currently, several ongoing projects need to evaluate assessor performance. One approach is the application of a true condition (= 'gold standard') to compare the individual assessor's scores with, and to subsequently calculate sensitivity and specificity of the sensory evaluation (Mathur et al., 2012), for illustration see Table 1. Here, the sensitivity means the probability of an assessor to detect truly boar tainted carcasses as tainted which is relevant for consumer acceptance. On the contrary, specificity refers to the ability of an assessor to correctly classify truly non-boar tainted carcasses as

[*] Corresponding author at: University of Göttingen, Department of Animal Sciences, Albrecht-Thaer-Weg 3, D-37075 Göttingen, Germany. Tel.: +49 551 39 5611; fax: +49 551 39 5587.
*E-mail address:* daniel.moerlein@agr.uni-goettingen.de (D. Mörlein).

http://dx.doi.org/10.1016/j.meatsci.2014.10.007
0309-1740/© 2014 Elsevier Ltd. All rights reserved.

**Table 1**
Confusion table to illustrate the relationship of test outcome and gold standard.

| | | **Condition** (as determined by "Gold standard") | | |
|---|---|---|---|---|
| | | positive* | negative | |
| **Test outcome** | *positive* | True positive (TP) | False positive (FP)('false alarm') | Positive predictive value (PPV) = TP/(TP + FP) |
| | *negative* | False negative (FN)('miss') | True negative (TN) | Negative predictive value (NPV) = TN/(TN + FN) |
| | | Sensitivity (SE) = TP/(TP + FN) | Specificity (SP) = TN/(TN + FP) | Accuracy (ACC) = (TP + TN)/(TP + TN + FN + FP) |

* 'positive' is referred to as boar tainted here while 'negative' means taint-free.

untainted which is related to logistics and extra costs. Compared to a composite indicator such as d′ (d prime) which is often used in discrimination testing (Brockhoff & Christensen, 2010), the estimates of sensitivity and specificity can easily be interpreted as percentages of "misses" (1 – sensitivity) and "false alarms" (1 - specificity). By contrast, d′ would give an overall estimate of the sensory difference of boar tainted vs. non-boar tainted samples in total.

The 'gold standard' could either be (i) chemical analysis while applying thresholds for androstenone and skatole, (ii) the sensory score of a reference assessor, (iii) average sensory score of a trained panel, or (iv) consumer acceptance scores using a threshold above which consumer liking is impaired. It is beyond the scope of this communication to discuss the advantages and drawbacks of these approaches. Note, however, that there is considerable variation between different protocols for chemical analyses (Ampuero Kragten et al., 2011) which in turn may affect the thresholds to be used when the chemical analysis is used to conclude on the 'true condition'. Furthermore, a recent study suggested that thresholds for skatole might be lower than previously established (Lunde, Skuterud, Hersleth, & Egelandsdal, 2010) while androstenone thresholds for impaired consumer acceptance could rather be raised in the absence of skatole (Bonneau & Chevillon, 2012; Meier-Dinkel, Trautmann, et al., 2013).

With this communication we aim (1) to illustrate how uncertainty about sensitivity and specificity is influenced by the total number of assessed samples and the prevalence of positive samples (=boar tainted samples), and (2) to provide easy-to-use figures, tables and computer code to obtain the necessary sample size to validate the sensitivity of a given assessor. The tools provided can be used by researchers and quality control officers to estimate the number of tainted carcasses that should be monitored to assess the efficacy of sensory evaluation to detect boar tainted carcasses. The sensitivity itself is regarded as the most relevant parameter from the perspective of consumer protection as it indicates the percentage of "misses"; sensitivity therefore needs to be considered in the sample size calculation. In what follows, all calculations were done using R (R Core Team, 2014).

## 2. Risk analysis (agreement of test outcome with true condition)

To analyse the agreement between sensory evaluation and chemical analysis, the absolute number and proportion of true negatives (TN), true positives (TP), false negatives (FN), and false positives (FP) is calculated. To do so, for example the chemical analysis can be used as "true condition" by applying thresholds above which carcasses are presumably tainted (e.g. skatole >0.2 μg/g and or androstenone >1.0 μg/g fat) to compare with the test outcome, i.e. sensory classification of a given assessor (sometimes using a post-hoc dichotomisation of the original sensory score). With respect to boar taint, TP refers to a truly boar tainted sample that is classified by an assessor as such. The individual score of an assessor's sensory evaluation is then compared to the true condition score resulting in TP, FP, FN, TN, respectively. Subsequently, sensitivity, specificity and accuracy are calculated according to, for example, Lalkhen and McCluskey (2008). Briefly, calculations were as follows: sensitivity SE = TP/(TP + FN), specificity SP = TN/(TN + FP). Accuracy is referred to as the proportion of correctly scores samples: ACC = (TP + TN)/(TP + TN + FN + FP). Furthermore, it has to be distinguished between the *observed* sensitivity as calculated above, and the assessor's true but latent sensitivity, which is the (unknown) true probability that a truly boar tainted carcass is detected.

## 3. Estimation of uncertainty for sensitivity (and specificity) under various scenarios

To illustrate the effect of sample size and prevalence on uncertainty for point estimates of sensitivity and specificity, several scenarios are compared for given parameters: (i) overall sample size (100, 1000); (ii) prevalence of tainted carcasses (10%, 20%, 50%) in the sample; and (iii) observed sensitivity (50%, 80%, 90%). Different types of confidence intervals for SE (and SP) were calculated: (a) the commonly used (textbook) asymptotic confidence interval using the normal approximation, (b) the approximate "Wilson" type interval (Wilson, 1927) propagated by (Agresti & Coull, 1998), and (c) the exact interval using the cumulative distribution function of the binomial distribution according to (Clopper & Pearson, 1934). Fig. 1 shows the corresponding 90% intervals for sensitivity. For calculation, we used the binconf() function from the R package Hmisc (Harrell, 2014).

We see that the uncertainty is huge when the sample size for sensitivity estimation is as low as 100 carcasses (Fig. 1). For example, the uncertainty for the point estimate of SE =80%, given a sample size of n =100 carcasses and prevalence =10%, ranges from 49.3% to 96.3% (exact method). After increasing the assumed prevalence from 10% to 50%, the confidence interval is considerably smaller (CI, 68.4% to 88.7%) as the number of tainted carcasses rises from 10 to 50, which lowers the variability of the estimate of sensitivity. For the same reason, when the sample size is increased to 1000 carcasses, the uncertainty for the point estimate of SE drastically decreases, even when the prevalence is as low as 10% (CI, 72.3% to 86.3%), and the interval is of course even shorter when the prevalence is 50% (CI, 76.8% to 82.8%). For comparison, industry estimates for prevalence of tainted carcasses are as low as about 4% (Mathur et al., 2012). To conclude, when evaluating the sensitivity of a given assessor, the sample size needs to be adjusted to the desired level of (un)certainty, given the estimated prevalence of tainted carcasses.

Concrete CIs as in Fig. 1 are available for observed sensitivities after making a validation experiment. When planning an experiment, by contrast, CIs are random (as the observed sensitivity is random) and we have to refer to the expected interval width given a true but latent sensitivity of the assessor. This information, however, can be used for two scenarios of performance validation: When (i) the validation is to be performed close to real-life situation in daily slaughter routine, e.g., when hot carcasses are tested in-line, one can only assume the prevalence of tainted carcasses from previous observations, e.g., from the same farm or breed type. Total sample size should therefore be chosen large enough based on assumed prevalence and with respect to the desired width of CI. On the contrary, it may (ii) be preferable to design a validation experiment, i.e. to present a certain number of tainted and untainted carcasses (as validated a priori) in random order. For example, in case (i), assuming a random sample of 1000 carcasses, in-sample prevalence of 5% tainted carcasses, and latent true SE = .80 using the Clopper & Pearson intervals results in an expected CI width of .23, while for case (ii) a planned experiment with 200 carcasses, in-