Food Control 125 (2021) 107940

Contents lists available at ScienceDirect

Food Control

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

Reliability of remote post-mortem veterinary meat inspections in pigs using augmented-reality live-stream video software

Viktor Almqvist *, Charlotte Berg, Jan Hultgren

Department of Animal Environment and Health, Swedish University of Agricultural Sciences, P.O. Box 234, 53223 Skara, Sweden

ARTICLE INFO

Keywords:
Abattoir
Augmented reality
Digital video communication
Remote guidance
Slaughter
Veterinary meat inspection

ABSTRACT

Official meat inspections at remotely situated, small-scale slaughter houses and game handlings establishments fare associated with a relatively high cost of official control per inspected animal. By performing veterinary meat inspections via live-streamed video, this cost could be lowered. We aimed to evaluate how veterinary meat inspections at slaughter can be conducted remotely with the help of a camera-equipped non-veterinary technician on site. Specialized software and augmented reality technology were used. The remote inspection was compared to standard on-site veterinary meat inspection at a large-scale slaughter plant for pigs in Sweden during 2019. The remote and on-site inspectors recorded findings in 400 carcasses and organs arrested for further inspection. The comparison was based primarily on percentage agreement, Cohen's kappa and prevalence- and bias-adjusted kappa (PABAK) as measures of *agreement* and *reliability*. The remote method was shown to display a high level of agreement for clear, easily distinguished findings (e.g. tail lesions, with an agreement of 92.3%, Cohen's kappa of 0.77 and PABAK of 0.85). For more vague findings and subjective decisions, the performance was slightly lower (e.g. whether or not to condemn a carcass completely, with agreement 75.2%, Cohen's kappa 0.32 and PABAK 0.50). Remote inspection appears to constitute a viable alternative for post-mortem meat inspection in pigs, given a sufficiently standardized method of inspection and sufficient inspection times. The performance of remote inspection probably depends on which persons use the method.

1. Introduction

At commercial slaughter in Europe, animals, carcasses and organs must be inspected, mainly to prevent and detect public health hazards (Council of the European Union, 2017). This is carried out by either Official Veterinarians (OV) or Official Auxiliaries (OA), in Sweden employed by the Swedish Food Agency. All animals are inspected ante-mortem (AM), and the carcasses and organs are inspected post-mortem (PM). AM inspections are performed by an OV, while PM inspections can be performed by either an OV or an OA by delegation.

PM inspections are carried out about halfway through the slaughter process, when the carcass has been gutted and split, but before trimming. The inspector checks for any signs of gross pathological lesions, or other issues related to food safety, contagious disease or animal welfare. In Sweden, findings are recorded using a system of two-digit codes for commonly occurring findings (Swedish Food Agency, 2012). If the carcass or organs display signs that may render them unfit for human consumption, they are marked, and will undergo a more thorough inspection by an OV. This routine is known as 'arresting for further inspection'. If the carcass or organs show signs of being a potential human or animal health hazard, they are declared unfit for human consumption by the OV and the entire animal is discarded (total condemnation, TC).

In 2014 a legislative shift was made in the EU to allow for a purely visual PM inspection of carcasses of pigs reared under so-called 'controlled conditions', essentially meaning indoors, instead of the previous palpation- and incision-based inspection (Council of the European Union, 2004; Hill et al., 2013; European Commission, 2014). This shift was motivated by a gradual change in the spectrum of hazards over the years, from gross pathological lesions to microbial contaminants. By reducing manual handling of the carcasses and organs, the risk of microbial contamination decreased. Furthermore, visual inspection improved cost-effectiveness because more carcasses could be inspected in the same time frame (Calvo-Artavia et al., 2013). Studies conducted beforehand (Mousing et al., 1997) as well as afterwards (Calvo-Artavia et al., 2013; Hill et al., 2013; Stärk et al., 2014), showed no markedly increased risk for the consumers as a result of this shift. Purely visual inspections were shown to be equally good at determining disease, but it

* Corresponding author.
E-mail addresses: viktor.almqvist@slu.se (V. Almqvist), lotta.berg@slu.se (C. Berg), jan.hultgren@slu.se (J. Hultgren).

https://doi.org/10.1016/j.foodcont.2021.107940
Received 25 November 2020; Received in revised form 24 January 2021; Accepted 26 January 2021
Available online 31 January 2021

was also concluded that a majority of registered findings do not pose a risk to the human population (Hill et al., 2013; Mousing et al., 1997).

Small-scale slaughter and game handling facilities contribute only 3.5% of the total Swedish red meat production, but still account for 26% of the total time spent on AM and PM inspections (Arja Helena Kautto, Swedish Food Agency, personal communication, October 22, 2020). These facilities operate without continuous manning by official control personnel, while larger facilities have control personnel present during all working hours. Control personnel, usually OVs, are required to visit these smaller plants once or twice per production day in order to perform the inspections. As the demand for locally produced meat increases, the number of small-scale plants is likely to increase with time.

With increasing fuel costs, climate awareness and concerns about the time control personnel spend driving to and from remote small-scale plants, inspection routines would benefit from modernization. In Sweden, the costs for manning slaughter plants with control personnel are mainly borne by the industry itself (Council of the European Union, 2017; Stärk et al., 2014). These costs have been described as "excessive" by the industry (Arzoomand et al., 2019).

The meat industry has long been quick to apply new technologies; from automated slaughter-line operation (Nielsen et al., 2014) to computer-vision aided meat quality assessment systems (Pabiou, 2012; Taheri-Garavand et al., 2019). These innovations are often motivated by economic gains, e.g. reduced staff or a reduced risk of human error, which can also affect the quality of the end product. The use of remote video transmission and video-assisted procedures is abundant in human medicine. A wide range of consultations within different fields can be performed remotely (Schroeder, 2019), and live video has also been used in different surgical procedures (Marescaux et al., 2002; Wang & Singh, 2017). Even the clinical veterinary sector has seen services emerge based on these ideas (Oxley & Saunders, 2015). Similar technologies and methodologies might be used for veterinary inspections at remote slaughter plants and game handling facilities, which could potentially lead to substantial financial and ecological benefits. To date, to our knowledge, there are no other scientific publications on the use of live video-based methods for veterinary meat inspection.

This study aimed to evaluate remote PM meat inspection at pig slaughter using two-way live-stream video communication with augmented reality software as an alternative to current on-site inspection practices. This was primarily accomplished by determining agreement and reliability between the two methods. The study was part of a project at the Swedish Food Agency to streamline and modernize public control at slaughterhouses and game handling facilities through, e.g., innovative and digital solutions.

2. Material and methods

2.1. Remote inspection setup

A comparative study was conducted during the spring of 2019 at a Swedish large-scale slaughter plant for pigs, processing around 3500 animals daily. A large plant was selected to achieve a sufficient sample size in a reasonable time period. Finishing pigs were typically slaughtered at 6 months of age, with a live weight of 120–130 kg.

Remote inspections were performed using a two-way remote video connection. On site at the slaughterhouse was a technician carrying a smartphone mounted on the back of the dominant hand and connected to a wireless headset. The smartphone was used to relay video and verbal information about the animals to the remote veterinarian, who could in turn give verbal commands to the technician. Furthermore, the remote veterinarian could employ an augmented reality overlay, directly showing his/her hand or other objects superimposed over the image the technician's smartphone display. This enabled the veterinarian to quickly show what the technician was expected to focus on, without complicated explanations.

The software solution used for remote inspection was XMReality Remote Guidance version 6.6.2 for Android and version 6.3.2–3 for Windows (XMReality AB, Linköping, Sweden). The transmitting device used was a Samsung Galaxy S9+ smartphone (Samsung Inc., Seoul, South Korea), running Android version 8.0.0 (Alphabet Inc., Mountain View, California, USA) and connected over WiFi to a local area network (LAN). The receiving terminal was a high-end, 8-core PC running Windows 10 Education version 1803 (Microsoft Corp., Redmond, Washington, USA) connected through Ethernet to the same LAN.

2.2. Sample

Finishing pigs were observed during normal slaughter on 27 days between March and June 2019. The sample consisted of 400 carcasses with associated red organs (heart, lungs, liver and kidneys and, when applicable, intestines) arrested by the OA for further inspection for any reason. Healthy carcasses and organs were also arrested and underwent the same inspection, thus forming a control group. The OA were instructed that these should be free of findings. Since healthy pigs dominated, healthy arrested carcasses were selected by the OA using a pre-generated random gamma-distributed list of 0–7 carcasses per work shift of approximately 1.5 h. The random process was designed to result in 50% healthy carcasses, in total, and the OA registered which carcasses (running number) these were. The study veterinarians and technician had no prior knowledge of which carcasses the OA had arrested despite being healthy. In case too many carcasses were arrested at one time (with insufficient time to inspect them all) some were excluded from the study, in order to not disturb the normal operations more than necessary. Carcasses that could not be inspected due to time constraints were skipped at random, so as not to introduce selection bias.

2.3. Data collection

All inspections were performed by two study veterinarians with several years of prior work experience in the field of meat inspection, and a technician without any experience of slaughter. Before the study, the veterinarians attended a half-day seminar on meat inspection organized by the Swedish Food Agency, and performed a two-day training session (inspection of 19 carcasses) together at the same facility, in an attempt to standardize their assessments. Neither veterinarian had performed meat inspection using a remote method prior to the study.

One of the study veterinarians performed on-site inspections (OSI), while the other worked with the camera-equipped technician to perform remote inspections (RI) (Fig. 1). The veterinarians switched roles three times during the trials, so that each would inspect 50% of the carcasses using either method. All inspections were carried out in an OSI-RI-OSI order. In this way, the OSI was divided into two parts, with the first part consisting of a visual inspection and the second being an assessment of incisions, if deemed necessary. Hence RI always started with untouched material without any prior manipulation which could otherwise have revealed the on-site inspector's suspicions.

The recordings were carried out according to a modified version of instructions of the Swedish Food Agency (2012). For each carcass, findings were registered using one or several of the 26 codes in Table 1, saved as binary variables (present or not). The codes represented common lesions or conditions, along with two classifications made by the veterinarians; false arrest (FA) and TC. FA indicated that the inspector judged that the carcass had been arrested as healthy. FA and TC classifications were based on the findings at inspection and were mutually exclusive (a carcass deemed falsely arrested could not be totally condemned, and vice versa). The absence of both FA and TC indicated that the carcass had been correctly arrested, but the findings were not severe enough to justify discarding the carcass.

Furthermore, the veterinarians scored their confidence in the findings and classification on a Likert-scale from 1 (not at all confident) to 5 (completely confident) for each carcass, and recorded the carcass'

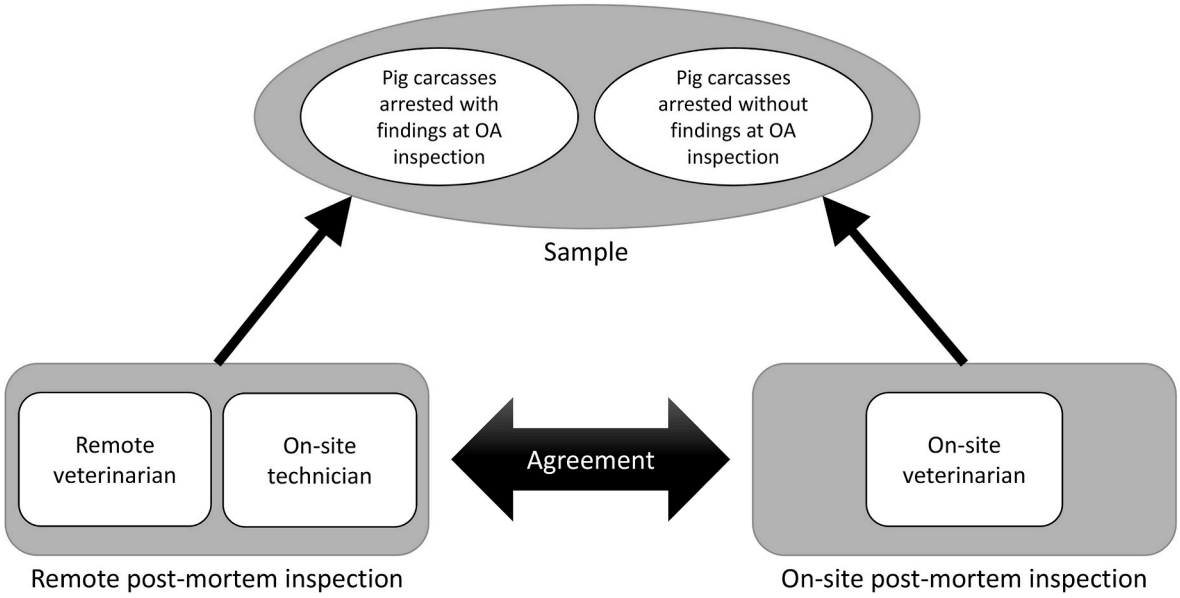

Fig. 1. Study design for comparing on-site PM inspection to remote PM inspection.

Table 1
List of codes for findings at inspections.

Code	Finding	Code	Finding	Code	Finding
06	Atypical mycobacteriosis	40	Old injury	72	Actinobacillus pleuropneumonia
18	Erysipelas	42	Recent injury	76	Pleuritis and/or endocarditis
19	Systemic infectious disease	46	Abnormal odour	78	Pleuritis and peritonitis
26	Tumour	48	Emaciation	84	Parasitic liver lesions, "white spots"
30	Abscess	52	Other finding[b]	88	Other liver lesions
32	Arthritis	56	Kidney lesion	999	No findings
34	Abnormal appearance	58	Tail lesion	FA	Falsely arrested
36	PSE[a]	62	Swine enzootic pneumonia	TC	Totally condemned
38	Fatty liver	64	Other pneumonia		

[a] Quality condition characterized by pale, soft and exudative meat.
[b] Used to denote conditions which have no code of their own. In this study predominantly splenic torsion.

production number, which was printed in 10-cm orange digits on its back.

Data were collected using an interactive macro-based form, written in Microsoft Visual Basic for Excel (Microsoft Corp., Redmond, Washington, USA), which emulated the touch-based terminals normally used for entering findings during meat inspection at Swedish slaughter plants. At onsite inspections, the form was used on a Windows 10-based tablet device (Microsoft Surface GO, Microsoft Corp., Redmond, Washington, USA), while the remote veterinarian used the same workstation as for the video inspections.

2.4. Statistical analysis

For each finding or classification, the prevalence, observed percentage agreement (joint probability of agreement), Cohen's kappa (Cohen, 1960), prevalence- and bias-adjusted kappa (PABAK; Byrt et al., 1993) and indices of prevalence and bias were calculated, as per recommendations laid out by Sim and Wright (2005), as well as associated 95% confidence intervals. Together, these values were intended to give a good picture of the method agreement, since no single value was judged to be sufficiently informative under these conditions. Prevalence was calculated as the average number of registrations for a certain finding between both veterinarians, divided by the number of inspected carcasses (400).

Calculations regarding total condemnation were based on the subset of carcasses where the veterinarians agreed that they were not FA, i.e. none of the veterinarians classified them as FA ($n = 153$). Average inspection time and average rater confidence score were calculated for each method, as well as the average change in confidence score with the number of carcasses inspected at both OSI and RI.

A comparison of recorded carcass production numbers between the methods was also performed, in which agreement was calculated as the number of carcasses where both veterinarians registered the same number divided by the number of carcasses (400). No further agreement statistics were calculated on these recordings.

Sensitivity and specificity estimates were based on the carcasses recorded by the veterinarians as FA and 'no finding' (code 999) respectively, evaluated against the list of FA carcasses produced by the OA (as gold standard), for OSI and RI separately. Thus, the sensitivity of FA at OSI was calculated as the proportion of all FA produced by the OA that the OSI veterinarian coded as FA, and the sensitivity of 'no finding' at OSI as the proportion of all healthy carcasses produced by the OA that the OSI veterinarian coded as 'no finding'.

All statistical calculations and analyses were performed in R (R Core Team, 2017). Cohen's kappa and PABAK were calculated using the function epi.kappa () in the package epiR (Stevenson & Reynard, 2020).

3. Results

Results from the data collection, together with agreement and kappa calculations, are presented in Table 2. Codes 18 (erysipelas), 38 (fatty liver) and 78 (pleuritis-peritonitis) were not recorded at all, and were excluded from subsequent analysis. A total of 220 (55%) of the carcasses were healthy. The most common findings were codes 999 (40.5%), 64 (22.8%) and 58 (21.1%), and the most uncommon were codes 46 (0.4%), 26 (0.8%), 40 (0.8%), 42 (0.8%) and 48 (0.8%). In total, 14 out of 23 codes had a prevalence of less than, or close to, 5%.

Table 2
Prevalence, inter-method reliability based on Cohen's kappa (95% confidence interval), prevalence- and bias-adjusted kappa (PABAK; 95% confidence interval) and prevalence and bias indices (95% confidence interval), as well as observed percentage agreement, for individual findings ($n = 400$), FA ($n = 400$) and TC ($n = 153$).

Code or classification	Prevalence, %	Cohen's kappa	PABAK	Prevalence index	Bias index	*Observed agreement, %*
06	1.3	0.59 (0.50–0.69)	0.98 (0.95–0.99)	0.98 (0.96–0.99)	0.00 (−0.02–0.02)	99.0
19	4.6	0.24 (0.14–0.33)	0.87 (0.81–0.91)	0.91 (0.88–0.94)	0.02 (−0.01–0.01)	93.3
26	0.38	0.66 (0.57–0.76)	0.99 (0.96–1.00)	0.99 (0.98–1.00)	0.00 (−0.01–0.01)	99.8
30	5.5	0.76 (0.66–0.86)	0.95 (0.91–0.98)	0.89 (0.86–0.92)	0.01 (−0.04–0.03)	97.5
32	4.2	0.57 (0.47–0.67)	0.93 (0.88–0.96)	0.92 (0.89–0.94)	0.01 (−0.02–0.04)	96.5
34	2.5	0.29 (0.22–0.36)	0.93 (0.88–0.96)	0.95 (0.93–0.97)	0.04 (0.010.06)	96.5
36	1.8	0.71 (0.61–0.81)	0.98 (0.95–0.99)	0.97 (0.95–0.98)	−0.01 (−0.02–0.01)	99.0
40	0.75	−0.01 (−0.11–0.09)	0.97 (0.94–0.99)	0.99 (0.97–1.00)	0.00 (−0.01–0.01)	98.5
42	0.75	0.33 (0.24–0.42)	0.98 (0.95–0.99)	0.99 (0.97–1.00)	0.01 (−0.01–0.02)	99.0
46	0.38	0.67 (0.57–0.76)	1.00 (0.97–1.00)	0.99 (0.98–1.00)	0.00 (−0.01–0.01)	99.8
48	0.75	0.33 (0.26–0.40)	0.98 (0.95–0.99)	0.99 (0.97–1.00)	0.01 (0.00–0.02)	99.0
52	5.0	0.74 (0.64–0.83)	0.95 (0.91–0.98)	0.90 (0.87–0.93)	0.01 (−0.02–0.04)	97.5
56	14.1	0.47 (0.38–0.57)	0.75 (0.67–0.81)	0.72 (0.67–0.76)	0.01 (−0.03–0.04)	87.3
58	21.1	0.77 (0.67–0.89)	0.85 (0.78–0.89)	0.58 (0.52–0.63)	0.01 (−0.02–0.04)	92.3
62	11.3	0.47 (0.38–0.57)	0.79 (0.72–0.85)	0.78 (0.73–0.82)	−0.02 (−0.06–0.02)	89.5
64	22.8	0.74 (0.65–0.87)	0.82 (0.76–0.87)	0.55 (0.49–0.60)	0.02 (−0.03–0.08)	91.0
72	1.1	0.21 (0.12–0.31)	0.97 (0.93–0.99)	0.98 (0.96–0.99)	0.00 (−0.01–0.02)	98.3
76	16.0	0.67 (0.56–0.76)	0.82 (0.76–0.87)	0.68 (0.63–0.73)	−0.02 (−0.07–0.03)	91.0
84	10.6	0.51 (0.41–0.61)	0.82 (0.75–0.89)	0.79 (0.75–0.83)	−0.01 (−0.06–0.03)	90.8
88	1.3	0.39 (0.29–0.49)	0.97 (0.94–0.99)	0.98 (0.96–0.99)	0.00 (−0.01–0.01)	98.5
999	40.5	0.63 (0.53–0.72)	0.64 (0.56–0.71)	0.19 (0.12–0.25)	0.01 (-0.01–0.02)	81.8
FA	58.4	0.86 (0.76–0.96)	0.87 (0.81–0.91)	−0.17 (−0.24–0.10)	0.03 (−0.08–0.06)	93.3
TC	23.5	0.32 (0.17–0.47)	0.50 (0.35–0.64)	0.53 (0.44–0.62)	0.11 (0.02–0.21)	75.2

Observed percentage agreement ranged between 75.2% (TC, $n = 153$) and 99.8% (codes 26, tumour and 46, abnormal odour, $n = 400$), with the majority of the values (19 of 23) above 90%, and the rest above 80%. Cohen's kappa ranged from −0.01 (code 40, old injury) to 0.86 (FA), with the majority of the values (12 of 23) between 0.2 and 0.6, and 8 values between 0.6 and 0.8. PABAK was between 0.50 (TC), and 1.00 (code 46, abnormal odour) with the majority of the values (19 of 23) above 0.8. Prevalence index ranged between −0.17 (FA) and 0.99 (codes 26, 40, 42, 46 and 48), with the majority of the values (14 of 23) above 0.85. Bias index ranged between 0.004 (code 46, abnormal odour) and 0.05 (code 999, healthy). These indices reflect the prevalence of findings and bias in the registrations, with values close to zero meaning a prevalence close to 50% and almost no bias. The distribution of findings did not differ significantly between OSI and RI, with the exception of code 34 (abnormal appearance) which was recorded 17 times at OSI compared to 3 times at RI, and TC with nearly twice as many carcasses condemned at OSI compared to RI (52 *vs.* 28). For FA, the difference was marginal (232 *vs.* 235 carcasses). The distribution of findings was similar in the subset of carcasses registered as FC by either method (16 *vs.* 2 carcasses with code 34 at OSI and RI, respectively).

The time per inspection ($n = 400$) was 113 ± 56 s at OSI and 340 ± 128 s at RI (mean ± SD). Thus, RI took, on average, 227 s longer to perform than OSI. The veterinarians scored the confidence in coding ($n = 400$) as 4.47 ± 0.96 points at OSI and 4.22 ± 0.78 points at RI (mean ± SD). Thus the average certainty was 0.25 points lower at RI. The confidence score increased on average by 0.0014 points per inspected carcass at RI, while the change at OSI was negligible.

The agreement of carcass production number was 97% ($n = 400$), and in all 12 non-agreeing observations, only one digit differed. Nine of these observations had either one missed digit, or one digit replaced by an adjacent one.

Sensitivity and specificity were generally lower for RI than OSI (Table 3). The veterinarians reported 232 and 235 FA at OSI and RI, respectively, and 165 and 160 of these carcasses were coded as having no findings. For FA, sensitivity was lower than specificity, and for 'no finding', the proportions were reversed, at both OSI and RI.

4. Discussion

Inter-rater agreement (inter-rater reliability) is the degree of agreement between different raters or judges. It can be estimated in a number of ways (Gwet, 2014). Different statistics are suitable in different situations. Percentage agreement (joint probability of agreement) is the proportion of times that the raters agree, and it is the simplest and least robust measure in a nominal rating system. Other statistics that have been proposed for nominal data, correcting for the fact that agreement may occur by chance, include Cohen's kappa (Cohen, 1960), Scott's pi (Scott, 1955), Fleiss' kappa (Fleiss, 1971) and prevalence- and bias-adjusted kappa (PABAK; Byrt et al., 1993). There is considerable controversy regarding the choice of statistic, and whether or not there is a need to correct for chance agreement (Uebersax, 1987). Still, Cohen's kappa is the most commonly used measure. Basically, the same statistics can be used to estimate the degree of agreement between different assessment methods, such as when comparing different diagnostic tests.

Interpretation of kappa values has been suggested as 0.01–0.20 representing *"None to slight agreement"*, 0.21–0.40 *"Fair agreement"*, 0.41–0.60 *"Moderate agreement"*, 0.61–0.80 *"Substantial agreement"* and 0.81–1.00 *"Almost perfect agreement"* (Landis & Koch, 1977). Negative values would instead indicate systematic disagreement. While being the de facto standard for studies of agreement, Cohen's kappa can be somewhat difficult to interpret (Di Eugenio & Glass, 2004; Sim & Wright, 2005) and the suggested grading can be somewhat arbitrary and rough (Landis & Koch, 1977; McHugh, 2012). Somewhat puzzling, identical levels of agreement can yield very different kappa values and, conversely, different levels of agreement may result in identical kappa values. It has been shown that Cohen's kappa tends to be lower at very low or very high prevalences, although agreement is the same (Byrt et al., 1993; Di Eugenio & Glass, 2004; Feinstein & Cicchetti, 1990;

Table 3
Sensitivity and specificity for false arrest (FA) and 'no finding' (code 999) at on-site and remote inspections ($n = 400$).

Coding	*Sensitivity on-site inspection*	*Specificity on-site inspection*	*Sensitivity remote inspection*	*Specificity remote inspection*
FA	0.90	0.97	0.86	0.95
No finding	0.94	0.70	0.90	0.65

Hallgren, 2012; Nelson & Edwards, 2008). The opposite situation occurs when one rater systematically overestimates a finding, so-called registration bias, which can increase kappa. Byrt et al. (1993) introduced the prevalence- and bias-adjusted kappa (PABAK) to deal with these two situations. It aims to estimate what Cohen's kappa would be, given the registrations, in a theoretical population with 50% prevalence and no registration bias. At 50% prevalence and no bias kappa and PABAK are identical. PABAK should be interpreted using the same levels as kappa.

The present study revealed a very high overall percentage agreement between the two methods, with a majority of values above 90%. The lowest value was seen for the perceived need for TC, with only 75.2% agreement. On the other hand, only one kappa value was above 0.8 (FA), with a further nine values above 0.6. Most of these poor-performing items had a prevalence of less than 5%. At first glance, considering kappa alone, this would indicate rather poor agreement between the methods. For PABAK, the results were very different; only four values were below 0.8, and only one of these was below 0.6 (TC). These discrepancies between the various statistics are most likely due to the large differences in prevalence between different items. The bias index was low throughout, with the exception of TC, suggesting that there was no serious over-registration of findings with either method. On the other hand, there seemed to be a systematic difference in both registrations of code 34 (abnormal appearance) and total condemnations, with RI classifying registering fewer code 34 and fewer carcasses as TC, which should overinflate the kappa values slightly. There was an almost perfect linear relationship between percentage agreement and PABAK.

Worth noting is that not all poor kappa values were associated with low prevalences. Six items with kappa far below 0.8 had a prevalence between 10 and 40%; codes 56 (kidney lesion), 62 (enzootic pneumonia), 76 (pleuritis/endocarditis), 84 (parasitic liver) and 999 (no findings), along with TC. The same items also displayed the lowest PABAK values, with four being less than 0.8 and the remaining two being close to 0.8. In our opinion, these types of findings can sometimes be rather vague (of either small size or slight colour change) or give room for subjective interpretation. The opposite also holds true; there were items with 5% prevalence or lower that yet performed fairly well when it came to kappa (above 0.74); codes 30 (abscess), 36 (PSE) and 52 (other finding). These findings are generally distinct and thus comparatively easy to spot and assess. The most challenging is arguably the assessment of TC, which is also the category that performed the worst overall (Cohen's kappa 0.32; PABAK 0.50). This poor performance could perhaps in part be attributed to the differences in registrations of code 34 (which can be grounds for TC). With 16 registrations at OSI, and only 2 at RI, this is a not insignificant portion of the total number of carcasses marked as TC. Code 34 is often used to denote colour differences, sometime subtle, in the carcass, and a possibility exists that due to technological limitations these did not translate well to the video stream used for OSI.

The best-performing assessment item was FA, which could also be considered clearly visible, since true arrests generally show obvious signs of disease. The discrepancies between FA and code 999 (which would ideally be the same) is probably due to small or vague findings on the carcass, missed by the OA but observed by one of the veterinarians. These findings were also missed by the other veterinarian, which is seen as poor agreement for code 999. Sensitivity at meat inspection in pigs has previously been shown to vary widely with organ system, from 0.16 for parasitic lesions to 0.92 for lesions in the respiratory system (Bonde et al., 2010). Poor sensitivity might account for minor changes that slipped through on healthy carcasses.

In this study, estimated sensitivity was lower than specificity for FA at both OSI and RI. This relationship is in line with trends shown by meat inspection personnel, as reported by Bonde et al. (2010), who showed varying sensitivity but near perfect specificity. In this study, the relationship was reversed for 'no finding', which may be due to the fact that these OA missed some findings. If not all 'healthy' carcasses were truly without lesions, it follows mathematically that the specificity may have been somewhat underestimated. Conversely, if the OA incorrectly arrested carcasses as 'not healthy', it would have led to a seriously inflated sensitivity. Thus, the reported sensitivities and specificities for 'no findings' seem credible under the conditions of the present study. Generally, RI produced slightly lower sensitivity and specificity estimates than OSI, which is not unreasonable, all things considered.

Another good agreement between the two methods was seen for carcass number, with a percentage agreement of 97.0%. In this case, the agreement was probably not subject to chance, in the sense that the veterinarian guessed a five-digit number correctly. In 9 out of 12 carcasses, the mismatches could be attributed to a missing digit or a digit being swapped to an adjacent one, most likely due to typing errors by the observer. Disregarding these errors, the overall agreement was 99.3%. The last three mismatches may have been due to illegible numbers on the carcasses.

For practical reasons, it was not possible to have the same veterinarian use both methods on the same carcasses. Therefore, the studied inter-method reliability was confounded by the inter-rater reliability, which was also unknown, albeit attempts were made to maximize it. Thus, it might be difficult to determine whether the results are due to the rater or the method. With the possibility of uneven sampling, it also cannot be assumed both raters observed the same distribution of findings using both methods. The observed agreement was expected to be good if and only if both reliabilities were high, and poor if either or both were low.

Logically, differences between veterinarians would have had the least impact on objective assessments, e.g. the presence of a disease with clear lesions, and a stronger effect on the more subjective classifications, such as TC. Poor agreement for more well-defined items could then have been due to either the method or the imperfect sensitivity or specificity of the OA, while poor agreement on more subjective findings more likely stemmed from inter-rater differences. However, due to the study design, this cannot be conclusively shown.

The prevalence of each finding was lower than expected for a population of purely arrested carcasses, due to the dilution with FA, and grossly inflated when compared to the population at the initial meat inspection level (OA inspection). Furthermore, the spectrum of findings was most likely shifted towards more "severe" lesions, since these are the ones that give cause for further inspection. Therefore, the reported Cohen's kappa values were probably lower then what could be expected from a sample of purely arrested carcasses, and much higher than if the sample was truly random from the entire population of slaughtered pigs. These relations are non-linear (Byrt et al., 1993), and as such difficult to predict. Based on the normal frequency of arrests at the facility we estimated the reported prevalence estimates to be between 50 and 100 times higher than what would be seen at the initial meat inspection level. Byrt et al. (1993) have previously argued against using kappa to compare reliability between different populations with varying prevalences, which makes it difficult to translate these kappa values to 'real' inspection situations. PABAK should however be unaffected by differences in the prevalence between sample and population, meaning it is easier to translate to a real-world scenario. The results should be translatable to a population of purely arrested animals, and would likely show higher reliability in such a setting.

Around half of the codes had a prevalence of less than 5%. At such a low prevalence, the question arises as to whether reliability can really be evaluated, despite potentially high values of agreement. Agreement could be high due to many negative cases where the veterinarians agreed, which might say very little about agreement on the very sparse positive cases. Consequently, results from findings with an extremely low prevalence should perhaps be taken with a grain of salt.

Feinstein and Cicchetti (1990) argued that the essence of kappa is to assume that each rater has a fixed prior probability of making positive or negative ratings, and that in a study population or sample with unknown distributions this would be an inappropriate assumption. The same authors argued that proportional agreement could be a sufficient metric in

a blinded study. Zhao et al. (2013) continued on this line of reasoning in stating that one of kappa's flaws is to always assume maximum randomness (which could lead to an underestimate of the agreement), and that while some random effects might be present they are most likely smaller, and the 'true' value of agreement probably lies somewhere in between kappa and the percentage agreement. We would therefore argue that all three presented metrics (percentage agreement, kappa and PABAK) should be viewed together, to illustrate the performance of the remote inspection method. With this in mind, the remote method seems to perform well, with most of the observed findings displaying high percentage agreement and PABAK, while kappa is generally lower. In most cases, this can be attributed to a low prevalence, and poor agreement is really only present for the more subjective classifications and vague findings. There seems to be a positive correlation between a finding's distinctiveness and the objectiveness of rating on the one hand, and the reliability between inspection methods on the other. In our opinion, this shows that the remote method could perform very well, assuming the inspected item, whether it is a pathological lesion or an identification number, is distinctly and clearly visible.

Hill et al. (2013) previously made a list of findings at slaughter based on the risk level for consumers. The authors claimed only two diseases as primarily infectious through the consumption of pork; round worms and acute pericarditis. Based on this claim, the observed differences in the methods' performance in this study would be of little importance for consumer safety. However, the differences may have consequences in other fields, such as meat quality control or presentations of animal health statistics.

The challenge with the remote method is to ensure that the veterinarian is actually shown all the existing changes, and is sufficiently skilled in evaluating the presented images. Most likely, this would require a systematic approach to the presentation of the material and more inspection time per carcass than on-site inspections. Löw et al. (2015) have previously noted that longer, more thorough video inspections leads to higher accuracy in human video diagnostics. As suggested by the discrepancies in proportions regarding code 34 there is also a need to ensure the used technology has a high enough colour accuracy. Further research in this area is needed, in order to establish a systematic approach suitable for use in a production setting, which balances accuracy and time usage. Notably, the method would need to be tested on different species and under different production conditions. In this study, the veterinarian had no prior experience using the method, and the technician was not used to slaughter and meat inspection. The increasing confidence with number of inspections suggests that the users became more familiar with the method over time. Therefore, results would probably improve further if the personnel underwent in-depth training using the remote method. The Identical Elements Theory proposed by Woodworth and Thorndike (1901) suggests that optimal transfer of training occurs when the training is performed in the same settings and context as the later work.

Meat inspection legislation has proven to be able to adapt to and meet the requirements of the industry, and has previously been changed without any substantial negative consequences for the consumers (Calvo-Artavia et al., 2013; Hill et al., 2013; Mousing et al., 1997; Stärk et al., 2014). The shift to a remote solution for PM meat inspection would, based on this initial research, be in line with the arguments for the previous change of methodology. Thus another shift in PM inspection methodology in the near future would not seem too farfetched.

5. Conclusions

The agreement between inspection using remote two-way video communication and on-site inspection at post-mortem meat inspection at pig slaughter is generally high, although it is difficult to directly translate all of the results to a real-world application. There appears to be a positive correlation between the distinctiveness of findings and the reliability between methods. All-in-all, remote inspection appears to constitute a viable alternative for post-mortem meat inspection in pigs, given a sufficiently standardized method of inspection and sufficient inspection times.

CRediT authorship contribution statement

Viktor Almqvist: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Charlotte Berg:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Jan Hultgren:** Conceptualization, Methodology, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

None.

Acknowledgements

This project was initiated and financed by the Swedish Food Agency. The authors thank the slaughterhouse for kindly allowing us to intrude on their activities, as well as Cecilia Wahlström, OV, and Tommy Karlsson, state inspector, for their invaluable participation in data collection. The Swedish Food Agency participated in planning and logistics.

References

Arzoomand, N., Vågsholm, I., Niskanen, R., Johansson, A., & Comin, A. (2019). Flexible distribution of tasks in meat inspection – a pilot study. *Food Control, 102*, 166–172. https://doi.org/10.1016/j.foodcont.2019.03.010

Bonde, M., Toft, N., Thomsen, P. T., & Sørensen, J. T. (2010). Evaluation of sensitivity and specificity of routine meat inspection of Danish slaughter pigs using Latent Class Analysis. *Preventive Veterinary Medicine, 94*, 165–169. https://doi.org/10.1016/j.prevetmed.2010.01.009

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*, 423–429. https://doi.org/10.1016/0895-4356(93)90018-v

Calvo-Artavia, F. F., Nielsen, L. R., & Alban, L. (2013). Epidemiologic and economic evaluation of risk-based meat inspection for bovine cysticercosis in Danish cattle. *Preventive Veterinary Medicine, 108*, 253–261. https://doi.org/10.1016/j.prevetmed.2012.11.002

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46. https://doi.org/10.1177/001316446002000104

Council of the European Union. (2004). Regulation EC No 854/2004 of the European Parliament and of the Council of 29 April 2004 laying down specific rules for the organisation of official controls on products of animal origin intended for human consumption. *Official Journal of the European Union, 139*, 206–320. http://eur-lex.europa.eu/.

Council of the European Union. (2017). Regulation EU 2017/625 of the European Parliament and of the Council of 15 March 2017 on official controls and other official activities performed to ensure the application of food and feed law, rules on animal health and welfare, plant health and plant protection products, amending Regulations (EC) No 999/2001, (EC) No 396/2005, (EC) No 1069/2009, (EC) No 1107/2009, (EU) No 1151/2012, (EU) No 652/2014, (EU) 2016/429 and (EU) 2016/2031 of the European Parliament and of the Council, Council Regulations (EC) No 1/2005 and (EC) No 1099/2009 and Council Directives 98/58/EC, 1999/74/EC, 2007/43/EC, 2008/119/EC and 2008/120/EC, and repealing Regulations (EC) No 854/2004 and (EC) No 882/2004 of the European Parliament and of the Council, Council Directives 89/608/EEC, 89/662/EEC, 90/425/EEC, 91/496/EEC, 96/23/EC, 96/93/EC and 97/78/EC and Council Decision 92/438/EEC (Official Controls Regulation)Text with EEA relevance. *Official Journal of the European Union, 95*, 1–142. http://eur-lex.europa.eu/.

Di Eugenio, B., & Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics, 30*, 95–101. https://doi.org/10.1162/089120104773633402

European Commission. (2014). Commission regulation (Ec) No 218/2014 of 7 March 2014 amending Annexes to Regulations (EC) No 853/2004 and (EC) No 854/2004 of the European Parliament and of the Council and Commission Regulation (EC) No 2074/2005. *Official Journal of the European Union, 69*, 95–98. http://eur-lex.europa.eu/.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378–382. https://doi.org/10.1037/h0031619

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed.). Gaithersburg, Maryland: Advanced Analytics, LLC.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*, 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hill, A., Brouwer, A., Donaldson, N., Lambton, S., Buncic, S., & Griffiths, I. (2013). A risk and benefit assessment for visual-only meat inspection of indoor and outdoor pigs in the United Kingdom. *Food Control, 30*, 255–264. https://doi.org/10.1016/j.foodcont.2012.04.031

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159. https://doi.org/10.2307/2529310

Löw, S., Erne, H., Schütz, A., Eingartner, C., & Spies, C. K. (2015). The required minimum length of video sequences for obtaining a reliable interobserver diagnosis in wrist arthroscopies. *Archives of Orthopaedic and Trauma Surgery, 135*, 1771–1777. https://doi.org/10.1007/s00402-015-2339-y

Marescaux, J., Leroy, J., Rubino, F., Smith, M., Vix, M., Simone, M., & Mutter, D. (2002). Transcontinental robot-assisted remote telesurgery: Feasibility and potential applications. *Annals of Surgery, 235*, 487–492.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*, 276–282.

Mousing, J., Kyrval, J., Jensen, T. K., Aalbæk, B., Buttenschnøn, B., Svensmark, P., & Willeberg, P. (1997). Meat safety consequences of implementing visual postmortem meat inspection procedures in Danish slaughter pigs. *The Veterinary Record, 140*, 472–477. https://doi.org/10.1136/vr.140.18.472

Nelson, K. P., & Edwards, D. (2008). On population-based measures of agreement for binary classifications. *Canadian Journal of Statistics, 36*, 411–426. https://doi.org/10.1002/cjs.5550360306

Nielsen, J. U., Madsen, N. T., & Clarke, R. (2014). Automation in the meat industry: Slaughter line operation. *Encyclopedia of Meat Sciences, 2*, 43–52.

Oxley, J., & Saunders, R. (2015). Potential for telemedicine. *Companion Animal, 20*, 702. https://doi.org/10.12968/coan.2015.20.12.702

Pabiou, T. (2012). Genetics of carcass composition in Irish cattle exploiting carcass video image analysis (Acta Universitatis Agriculturae Sueciae, 2012:5). [Doctoral dissertation, Swedish University of Agricultural Sciences]. SLU Epsilon Open Archive https://pub.epsilon.slu.se/8533/.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Schroeder, C. (2019). Pilot study of telemedicine for the initial evaluation of general surgery patients in the clinic and hospitalized settings. *Surgery Open Science, 1*, 97–99. https://doi.org/10.1016/j.sopen.2019.06.005

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321. https://doi.org/10.1086/266577

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*, 257–268. https://doi.org/10.1093/ptj/85.3.257

Stärk, K. D. C., Alonso, S., Dadios, N., Dupuy, C., Ellerbroek, L., Georgiev, M., Hardstaff, J., Huneau-Salaün, A., Laugier, C., Mateus, A., Nigsch, A., Afonso, A., & Lindberg, A. (2014). Strengths and weaknesses of meat inspection as a contribution to animal health and welfare surveillance. *Food Control, 39*, 154–162. https://doi.org/10.1016/j.foodcont.2013.11.009

Stevenson, M., & Reynard, C. (2020). epiR: Tools for the analysis of epidemiological data. Melbourne, Australia https://cran.r-project.org/web/packages/epiR/epiR.pdf.

Swedish Food Agency. (2012). Kontroller vid slakt [Controls at slaughter]. https://www.livsmedelsverket.se/produktion-handel–kontroll/livsmedelskontroll/offentlig-kontroll/kontroller-vid-slakt Accessed October 4 2020.

Taheri-Garavand, A., Fatahi, S., Omid, M., & Makino, Y. (2019). Meat quality evaluation based on computer vision technique: A review. *Meat Science, 156*, 183–195. https://doi.org/10.1016/j.meatsci.2019.06.002

Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin, 101*, 140–146. https://doi.org/10.1037/0033-2909.101.1.140

Wang, S. C., & Singh, T. P. (2017). Robotic repair of a large abdominal intercostal hernia: A case report and review of literature. *Journal of Robotic Surgery, 11*, 271–274. https://doi.org/10.1007/s11701-017-0675-3

Woodworth, R. S., & Thorndike, E. L. (1901). The influence of improvement in one mental function upon the efficiency of other functions. (I). *Psychological Review, 8*, 247–261. https://doi.org/10.1037/h0074898

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association, 36*, 419–480. https://doi.org/10.1080/23808985.2013.11679142