Probability Assessments from Multiple Experts: Qualitative Information is More Robust

Linda C. van der Gaag [a] Silja Renooij [a] Hermi J.M. Schijf [a]

Armin R. Elbers [b] Willie L. Loeffen [b]

[a] *Department of Information and Computing Sciences, Utrecht University, NL*
[b] *Department of Virology, Central Veterinary Institute of Wageningen UR, NL*

Abstract

For many application domains, Bayesian networks are designed in collaboration with a single expert from a single institute. Since a network is often intended for wider use, its engineering involves verifying whether it appropriately reflects expert knowledge from other institutes. Upon engineering a network intended for use across Europe, we compared the original probability assessments obtained from our Dutch expert with assessments from 38 experts in six countries. While we found large variances among the assessments per probability, very high consistency was found for the qualitative properties embedded in the series of assessments per assessor. The apparent robustness of these properties suggests the importance of capturing them in a Bayesian network under construction.

1 Introduction

Bayesian networks are rapidly becoming the models of choice for reasoning with uncertainty in decision-support systems, most notably in application domains governed by physical, biological or chemical processes. While much attention has focused on algorithms for learning Bayesian networks from data, our experiences in designing networks for the biomedical field show that systematically collected data are often wanting or are not amenable to automated model construction. Often therefore, expert knowledge constitutes the only source of information for a network's design. Since the construction of a high-quality Bayesian network is a difficult and time-consuming creative process, both for the engineers involved and for the consulted experts, common engineering practice is to closely collaborate with just a single, or a very small number of experts, even if the network is intended for much wider use.

In collaboration with Dutch experts, we are in the process of developing a decision-support system to supply veterinary practitioners with an additional tool for the early detection of Classical Swine Fever (CSF) in pigs. At the core of the system lies a Bayesian network for computing the posterior probability of a CSF infection being present, given the clinical signs observed at a pig farm by an attending veterinarian. The network is being constructed with the help of an experimental CSF expert and a senior epidemiologist from the Central Veterinary Institute in the Netherlands. For its design, in-depth interviews were held with the two participating experts and case reviews were conducted with eight Dutch swine practitioners, both with and without clinical CSF experience. The conditional probabilities required for the network were mostly not available from the literature, nor were sufficiently rich data available for their estimation. As a consequence, all required probabilities were assessed by a single CSF expert.

While being built with Dutch experts, our Bayesian network for the early detection of Classical Swine Fever is intended for use across the European Union. A Bayesian network, in fact, is often intended for wider use than just by the experts with whom it is being constructed. Engineering a network then involves verifying whether it appropriately reflects practices and insights from other experts as well. Upon engineering our CSF network, we had the opportunity of attending project meetings with 38 pig experts and veterinary practitioners in six European countries outside the Netherlands. During these meetings, we were granted time to discuss with the experts some details of the current network and its detection abilities. Among other

information, we gathered assessments for a limited number of conditional probabilities for our network. Our intention was not to elicit probability assessments from multiple experts in order to aggregate these for use in our network. Rather, we were interested in whether or not experts from different countries would provide similar assessments for relations between diseases and clinical signs that were supposed to hold universally across countries. We therefore compared the obtained assessments with each other and with the original assessments provided by our Dutch expert.

During the project meetings, we obtained a total of 58 series of probability assessments from 38 domain experts in six countries. We investigated the assessments obtained for the separate probabilities by establishing various summary statistics, both per country and across countries. We further studied the series of assessments obtained per expert and the qualitative properties of dominance embedded in these series. We found large variances among the numerical assessments per probability, both within and between countries. Much higher consistency was found for the dominance properties embedded in the series of assessments, however. Apparently, this qualitative information is more robust than the numerical probability assessments themselves. This robustness suggests the importance of explicitly eliciting qualitative properties of probability and ensuring that these are properly captured in a Bayesian network under construction.

This paper reports on our findings and experiences from the project meetings, and is organised as follows. In Section 2, we briefly introduce the background of our application. Section 3 describes the set-up of the meetings and the elicitation method used. Section 4 summarises the numerical assessments obtained. In Section 5, we analyse our findings in qualitative terms. The paper ends with our reflections in Section 6.

2 The Context

In the context of a European project involving seven countries, a decision-support system is being developed for the early detection of Classical Swine Fever in pigs. CSF is a highly infectious viral disease of pigs that has a potential for rapid spread. The virus causing the disease is transmitted mainly by direct contact between infected and non-infected susceptible pigs. When a pig is first infected with the virus, it will show an increased body temperature and a sense of malaise, associated with such clinical signs as a lack of appetite and lethargy. Later in the infection, the animal is likely to develop an inflammation of the intestinal tract; also problems with the respiratory tract are beginning to reveal themselves through such signs as a conjunctivitis, snivelling, and coughing. The final stages of the disease are associated with an accumulating failure of body systems, which will ultimately cause the pig to die [1]. The longer a CSF infection remains undetected, the longer the virus can circulate without hindrance, not just within a herd but also between herds, with major socio-economic consequences. Yet, the aspecificity of especially the early signs of the disease causes the clinical diagnosis of CSF to be highly uncertain for a relatively long period after the infection has occurred.

Within the European CSF project, we are developing a decision-support system that supplies veterinary practitioners with an additional independent tool to identify CSF-suspect situations as early on in an outbreak as possible. The system takes for its input the clinical signs seen at a pig farm by an attending veterinarian and computes the probability of a CSF infection being present; based upon the computed probability, a recommendation for further proceedings is given. For computing the posterior probability of CSF given the observed clinical signs, the system builds upon a Bayesian network which models the pathogenesis of the disease. Figure 1(a) shows the network's graphical structure; it currently includes 32 stochastic variables, for which over 1100 (conditional) probabilities are specified.

3 Set-up of the Project Meetings

Between December 2006 and May 2007, project meetings were held with veterinary experts in Belgium, Denmark, Germany, Great-Britain, Italy, and Poland. Each meeting was organised in a renowned institute in the country being visited. For the meeting, a small number of veterinary experts from all over the country were invited; the invitees ranged from veterinary pig practitioners to researchers conducting experimental CSF studies. During these meetings, we were granted some time to discuss details of the CSF network and its detection abilities. Within the alotted time, in all countries, the experts were presented with the same lecture about the working of the network; in addition, the assessment task to be performed was introduced.

For the assessment task, a tailored elicitation method was used. The basic idea of this method is to present each requested probability to the assessor as a fragment of text stated in veterinary terms instead of in mathematical notation; the fragment is accompanied by a vertical scale with numerical and verbal

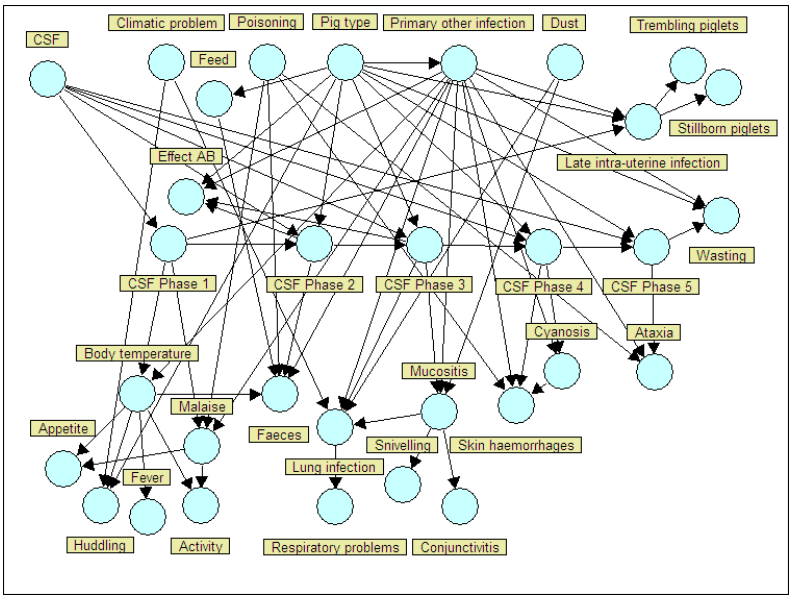

Figure 1: The graphical structure of the Bayesian network for the early detection of CSF

anchors. Figure 2 shows an example fragment of text, along with the probability scale. The assessor is asked to carefully consider the fragment of text and to indicate his assessment for the requested probability by marking the scale. For further details of the elicitation method, we refer the reader to [3]. The use of the probability scale was demonstrated in each meeting during the plenary instruction for the assessment task.

For our investigations, we selected twelve probabilities. In the present paper, we focus on six of these; for the other six probabilities similar results were found. The six probabilities under study in this paper are summarised in Table 1 and were elicited from the experts in the displayed order. The probabilities p_1 through p_4 denote the probabilities of finding the typical tear marks associated with a conjunctivitis (abbreviated to 'conjunct'), in an animal in the early stages of a CSF infection ('csf') and, respectively, no further primary infections ('no-other'), a respiratory infection ('resp'), a gastro-intestinal infection ('intest'), and both types of primary infection ('resp+intest'); note that in the current network the variable *Conjunctivitis* is related indirectly to both *CSF* and *Primary other infection*. The probabilities p_5 and p_6 denote the probabilities of finding the clinical sign of snivelling ('sniv') in an animal with or without an infection of the mucous in the upper respiratory tract, respectively; these two probabilities completely specify the conditional probability table for the variable *Snivelling* in the network. For comparison purposes, Table 1 further includes the original assessments provided by our Dutch expert during the initial elicitations for the network's construction.

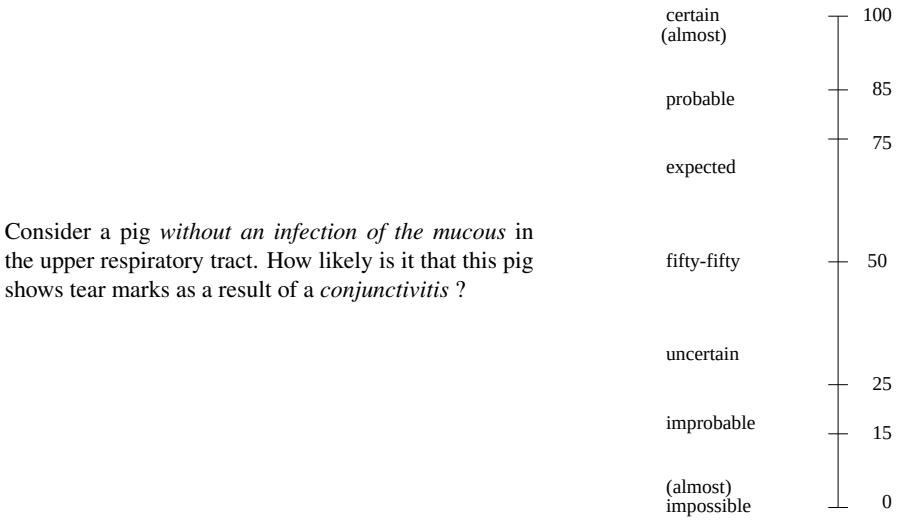

Figure 2: A fragment of text describing a requested probability, and the accompanying probability scale.

Table 1: The six probabilities discussed in this paper, with the assessments provided by the Dutch expert.

Probability	*Initial assessment*
$p_1 = \Pr(\text{conjunct} \mid \text{csf, no-other})$	0.29
$p_2 = \Pr(\text{conjunct} \mid \text{csf, resp})$	0.66
$p_3 = \Pr(\text{conjunct} \mid \text{csf, intest})$	0.29
$p_4 = \Pr(\text{conjunct} \mid \text{csf, resp+intest})$	0.66
$p_5 = \Pr(\text{sniv} \mid \text{muco})$	0.20
$p_6 = \Pr(\text{sniv} \mid \text{no-muco})$	0.01

With the set-up outlined above, we obtained assessments for the conditional probabilities p_1 through p_6 from a total of 38 veterinary experts in six countries. In the sequel, we will refer to these countries by the letters $\mathcal{A}$ through $\mathcal{F}$, for reasons of anonymity.

4 A Quantitative Analysis

We investigated the separate assessments obtained by establishing various summary statistics, both per country and across countries. In this section, we report these statistics and review our findings.

4.1 The Data Obtained, the Analyses and the Results

Upon investigating the responses obtained from our elicitation efforts, we noticed that the veterinary experts had used different methods for indicating their assessments on the probability scales. Most experts had put an explicit mark on the vertical line of the scale, as was demonstrated during the plenary instruction. The positions of these marks were measured and translated into numerical probability assessments for further analysis. Some experts, however, had encircled one of the verbal anchors positioned beside the scale. Since these anchors indicated a more or less fuzzy probability range [7], these circles were not used for numerical analysis. We obtained 58 completely specified series of assessments from our 38 experts: 29 series for the probabilities p_1 through p_4, and 29 series for the probabilities p_5 and p_6. In incomplete series, another 10 assessments were given, providing us with a total of 184 numerical assessments.

For each probability under study, we computed standard statistics over the various numerical assessments obtained. More specifically, we established the range, mean and standard deviation of the assessments per country; we further determined the mean and standard deviation of the six country means. Table 2 shows the resulting statistics for the probability p_1 in some detail; the statistics for the remaining five probabilities are provided in Table 3. We further computed some statistics per probability over all countries. The results are summarised in Table 4; note that the overall mean per probability may differ from the mean of the country means as a result from unequal sizes of the groups of assessors per country.

To conclude, we tested the null hypothesis of equal country means for each probability under study. For this purpose, we performed an analysis of variance with the F-statistic using a significance level of 0.05. For all probabilities except p_5, the null hypothesis of equal means across countries was rejected. For the probabilities p_1 through p_4 and p_6, we further performed post-hoc testing of pairwise equality under the

Table 2: The ranges, means $\overline{x}$ and standard deviations s of the assessments obtained for the probability p_1 under study, per country; assessments and means in bold fall within the modal interval [0.7,0.8).

Country	n	*Assessments*							*Range*	$\overline{x}$ (s)
$\mathcal{A}$	5	0.60	**0.75**	**0.75**	**0.75**	0.80			[0.60, 0.80]	**0.73** (0.08)
$\mathcal{B}$	6	0.30	0.40	0.50	**0.71**	**0.75**	0.85		[0.30, 0.85]	0.59 (0.22)
$\mathcal{C}$	5	0.15	0.15	0.20	0.25	0.30			[0.15, 0.30]	0.21 (0.07)
$\mathcal{D}$	5	0.40	0.50	**0.75**	0.90	0.95			[0.40, 0.95]	**0.70** (0.24)
$\mathcal{E}$	3	**0.70**	**0.75**	**0.79**					[0.70, 0.79]	**0.75** (0.05)
$\mathcal{F}$	7	0.15	0.34	0.50	0.64	**0.75**	**0.75**	**0.79**	[0.15, 0.79]	0.56 (0.24)
All means									[0.21, 0.75]	0.59 (0.20)

Table 3: The means $\overline{x}$ and standard deviations s of the assessments obtained for the probabilities p_2, p_3, p_4, p_5 and p_6 per country; means in bold fall within the relevant modal interval.

	p_2		p_3		p_4		p_5		p_6	
Country	n	$\overline{x}$ (s)	n	$\overline{x}$ (s)	n	$\overline{x}$ (s)	n	$\overline{x}$ (s)	n	$\overline{x}$ (s)
$\mathcal{A}$	5	**0.80** (0.08)	5	**0.70** (0.09)	5	**0.81** (0.08)	5	0.58 (0.22)	5	**0.15** (0.06)
$\mathcal{B}$	6	0.77 (0.18)	6	0.58 (0.21)	6	**0.82** (0.11)	7	**0.78** (0.20)	6	0.47 (0.28)
$\mathcal{C}$	6	0.27 (0.31)	5	0.24 (0.08)	6	0.43 (0.25)	6	0.68 (0.19)	5	**0.13** (0.06)
$\mathcal{D}$	5	0.70 (0.22)	4	0.46 (0.30)	4	0.78 (0.21)	3	0.82 (0.06)	3	0.50 (0.35)
$\mathcal{E}$	3	0.78 (0.08)	3	**0.74** (0.06)	2	**0.82** (0.04)	3	0.83 (0.20)	4	0.46 (0.38)
$\mathcal{F}$	7	0.75 (0.15)	7	0.65 (0.17)	7	0.75 (0.15)	7	**0.78** (0.05)	7	**0.19** (0.06)
All means	6	0.68 (0.21)	6	0.56 (0.19)	6	0.73 (0.15)	6	**0.75** (0.10)	6	0.32 (0.18)

assumption of equal variances. Post-hoc testing for the probability p_6 did not reveal any significant pairwise differences of the means per country. For the probabilities p_1 through p_4, however, post-hoc testing showed significant pairwise differences involving country $\mathcal{C}$. More specifically, for the probability p_1, $\mathcal{C}$'s country mean was found to be different from the country means of both country $\mathcal{A}$ and country $\mathcal{E}$. For the probability p_2, $\mathcal{C}$'s country mean differed from the country means of each of the other countries. $\mathcal{C}$'s country mean for p_3 differed from those of countries $\mathcal{A}$, $\mathcal{E}$ and $\mathcal{F}$. For probability p_4, to conclude, $\mathcal{C}$'s country mean was different from the country means of both $\mathcal{A}$ and $\mathcal{B}$. No further significant differences were found.

4.2 Discussion

The results of the numerical analyses per probability show very little consensus in the assessments obtained per country and across countries. Of the studied statistics, the most robust information appears to be provided by the modal interval per probability, which includes between 27% and 47% of the assessments. Yet, these modal intervals include only few of the reported means and none of the assessments provided by our Dutch expert. More specifically, only the overall mean of the assessments for the probability p_5 lies in the relevant modal interval; the overall means found for the probabilities p_1 through p_4 all lie in lower-ordered intervals, while the mean of the assessments for p_6 lies in a higher-ordered interval than the modal one. With respect to the country means, we further find for the probabilities p_1 through p_4 that those from country $\mathcal{A}$ all lie in their modal intervals; the country means for the probabilities p_1, p_3 and p_4 from country $\mathcal{E}$ also lie in their respective modal intervals; for country $\mathcal{D}$, this observation applies only to the mean for the probability p_1; and, for country $\mathcal{B}$, it holds for just the mean for p_4. All other country means for p_1 through p_4 lie in lower-ordered intervals. For the country means for the probabilities p_5 and p_6, we observe that those from countries $\mathcal{B}$ and $\mathcal{F}$ lie in their respective modal intervals; for p_6 this also holds for the mean from country $\mathcal{C}$. For the probability p_5, the country means computed from $\mathcal{A}$'s and $\mathcal{C}$'s assessments lie in lower-ordered intervals; all other country means for p_5 and p_6 lie in higher-ordered intervals.

Since the elicitation efforts in the six countries were not conducted in a controlled laboratory setting, numerous factors may have influenced the assessments, ranging from the way the task was introduced, through language barriers, the attitudes of the national experts and their levels of expertise, to the atmosphere in the group. Among all these factors, a likely explanation for the large differences in numerical assessments obtained is found in the varying levels and expertise of the assessors, even so within the focused area of Classical Swine Fever. According to the theory of naive probability [5], people can reason correctly about

Table 4: The ranges, modal intervals *mod* with frequencies #, means $\overline{x}$, and standard deviations s of all assessments obtained per probability; means in bold fall within the relevant modal interval.

	n	*range*	*mod* (#)	$\overline{x}$ (s)
p_1	31	$[0.15, 0.95]$	[0.7,0.8) (12)	0.58 (0.25)
p_2	32	$[0.10, 1.00]$	[0.8, 0.9) (10)	0.67 (0.27)
p_3	30	$[0.15, 0.85]$	[0.7, 0.8) (8)	0.56 (0.23)
p_4	30	$[0.20, 1.00]$	[0.8, 0.9) (10)	0.72 (0.21)
p_5	31	$[0.26, 1.00]$	[0.7, 0.8) (12)	**0.74** (0.18)
p_6	30	$[0.05, 0.96]$	[0.1, 0.2) (14)	0.30 (0.25)

probabilities by mentally considering and numbering the various possibilities in which an event may or may not occur. As such, probability estimates are highly influenced by the experience of the assessor, which would explain the variation in assessments provided by the experts in our investigations. An interesting finding in this respect is that in some countries the first assessments were rather close to one another, while in other countries larger ranges were found; this closeness of assessments may be explained by similarities in background and experience, yet may also have arisen from a bias introduced by someone remarking out loud that, for example, some scenario is quite likely. Another explanation for the observed differences lies in a heuristic called *anchoring and adjusting*, which is commonly found in people when asked to provide an assessment for a probability: using this heuristic, they choose a relevant known probability as an anchor to tie their assessment to by adjustment. From cognitive-science studies, it is well known that even for self-generated anchors, the adjustments made are typically insufficient [2, 4]. Since our assessors generated the first assessment in each series by consulting their memory, variations in these first assessments inevitably caused variations in the subsequent related assessments by the anchoring-and-adjusting heuristic.

While the referenced theories can explain the variation among assessors, they do not explain the observed differences between countries. Remarkable differences were found, for example, for the country means for each of the probabilities p_1 through p_4 established from the assessments from country $\mathcal{C}$, compared to the country means from the other countries. A possible explanation is that the experts from country $\mathcal{C}$ found the four probabilities very hard to assess, because these were conditioned on the presence of a CSF infection and, as they stated, "we don't have CSF in our country". Another possible explanation, supported by the sound recording of the elicitation, is that the experts actually assessed the complements of the requested probabilities: during the meeting, a moderator had translated the fragments of text into the experts' mother tongue and, from yet another native speaker who listened to the recording, we got strong indications that the translations were not entirely to the point. A third, less likely, explanation is that the experts from country $\mathcal{C}$ showed other biases than the assessors from the other countries.

Differences were also found between the assessments provided by our Dutch expert and the assessments obtained from the experts from the other countries: the Dutch assessments all lie in lower-ordered intervals than the respective modal intervals derived from all other assessments. An explanation for this finding may be that our expert provided his assessments from an entirely different background from the other assessors: the Dutch expert had been closely involved in the construction of the network for more than two years and had provided all probabilities required for its quantification, while the other assessors did not have intimate knowledge of the network and were confronted with a few probabilities in a single day's meeting. Moreover, as a result of the one-on-one elicitation sessions with our Dutch expert, any questions regarding a requested probability could be answered and obvious errors or inconsistencies could be prevented. In addition, our expert was explicitly trained in treating any variable not mentioned in a requested probability, as an unknown. Although this issue was elaborated upon in the plenary instruction for the other experts, it is not unlikely that probabilities were assessed in the context of a default value for unmentioned variables. Yet, the observed differences in assessments between the Netherlands and the other countries could also be due to entirely different factors, such as actual differences in housing, in feed and in health conditions. Before any definite conclusions can be drawn, therefore, we should obtain more insight in the differences, for example by repeating the study with a group of Dutch veterinary experts or by tailored elicitation from veterinary experts from various European countries in a more controlled setting.

5 A Qualitative Analysis

In the previous section, we reviewed numerical properties of the probability assessments obtained from the veterinary experts in the six visited countries. From our investigations, we concluded that the assessments showed little consensus. We now turn to a qualitative analysis of the assessments obtained, in which we consider qualitative properties embedded in the series of assessments provided per expert.

5.1 The Data Obtained, the Analyses and the Results

For our qualitative analysis, we had available the same 58 completely specified series of numerical assessments from which we established standard statistics in the previous section. In addition to these numerical series, we had also obtained, during the project meetings, 10 complete sets of verbal assessments, that is, assessments composed of encircled verbal anchors from the probability scale. Because the anchors indicated a fuzzy probability range, these assessments were not used in our quantitative analysis. We can now include

them in a qualitative analysis, however, because the rank order of the scale's anchors can be considered stable [7]. For our qualitative analysis, we observe that although the six probabilities under study are probabilistically independent, they are not so from a domain point of view: the probabilities p_1 through p_4 are related, as are the probabilities p_5 and p_6. Based upon common knowledge, for example, we can state that a pig with a mucositis in the upper respiratory tract is more likely to snivel than a pig without a mucositis. This statement essentially expresses that more severe clinical signs are more likely given more severe values on a disease scale. Properties stating that one conditional probability distribution is ranked as superior to another, are called properties of dominance [6]. In this section, we investigate the dominance properties embedded in the series of assessments provided by the 38 experts.

For studying properties of dominance, a total ordering of the conditioning contexts in the series of probabilities under study is required. For the probabilities p_1 through p_4 for this purpose a total ordering of the possible other primary infections is needed; based upon domain knowledge, we decided to use the ordering 'no-other' $<$ 'intest' $<$ 'resp' $<$ 'resp+intest'. For the probabilities p_5 and p_6, we chose the ordering 'no-muco' $<$ 'muco' for their conditioning contexts. For studying dominance properties, also a total ordering on the probabilities themselves is required. For the numerical assessments, the standard numerical ordering is used. For the verbal assessments, we assumed the ordering on the verbal labels as dictated by the probability scale, that is, we assume 'impossible' $<$ 'improbable' $< \ldots <$ 'probable' $<$ 'certain'. Based upon common knowledge, we should now find the following dominance properties in the series of assessments:

- $p_1 \leq p_3 \leq p_2 \leq p_4$;
- $p_6 \leq p_5$.

We would like to note that the original assessments from our Dutch expert indeed exhibit these properties.

For the probabilities p_1 through p_4, the assessments of 18 of the 29 experts (62%) who gave a complete numerical series, were found to obey the expected dominance property. Violations of the property were mostly found in assessments from experts from the countries $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$. In seven series, the violation was caused by the assessment for the probability p_1 being too high compared to that for either p_2, p_3 or p_4; in the other violating series, the assessment for the probability p_4 was too low compared to that for p_2. The assessments of three of the five experts (60%) who gave a complete set of verbal assessments for p_1 through p_4, also obeyed the expected dominance property. For the probabilities p_5 and p_6, we found that the assessments of 28 of the 29 experts (97%) who gave a complete numerical series, exhibited the expected property of dominance. The only violation was caused by the assessments $p_5 = 0.40$ and $p_6 = 0.50$, given by an expert from country $\mathcal{B}$. The assessments of all five experts (100%) who gave a complete set of verbal assessments for p_5 and p_6, embedded the expected dominance property.

5.2 Discussion

The results of our qualitative analysis show that the dominance properties embedded in the series of assessments obtained are far more consistent among the individual experts and across countries, than the quantitative properties studied in Section 4. With respect to the probabilities p_1 through p_4, for example, a relatively large number of experts (62%) expressed the expected property of dominance by providing assessments with $p_1 \leq p_3 \leq p_2 \leq p_4$. This finding is of interest since the probabilities were presented to the experts for assessment in a different order: the assessors thus did not simply provide increasingly higher, or lower, values. Assuming that they employed an anchoring-and-adjusting heuristic, this finding means that after providing an assessment for p_1, an assessor adjusted towards a higher value for p_2; for the probability p_3, he subsequently adjusted to a lower value, yet not below his initial assessment for p_1; for the final probability in the series, again an adjustment towards higher values was performed, to beyond the assessment for p_2. Of further interest is the finding that six violations of the property of dominance among the probabilities p_1 through p_4 were caused not by an adjustment in the wrong direction, that is, wrong in terms of the total ordering assumed, but rather by a wrong amount of adjustment. More specifically, after having provided an assessment for p_2, the adjustment to a lower value for p_3 was too large, with p_3 ending up smaller than p_1; alternatively, after having provided an assessment for p_3, the adjustment to a higher value for p_4 was not large enough, with p_4 ending up smaller than p_2.

To the best of our knowledge, researchers have not addressed anchoring and adjusting in tasks where subjects assess a series of more than two related probabilities. It is unknown, therefore, whether people would typically use the first anchor for all subsequent assessments, or tie each assessment to the previous one. The only mention of relating assessments for different conditional distributions was in the concept of

trend [3], which was spontaneously provided by assessors and explicitly revealed the anchor and amount of adjustment used. Our elicitation task for the two probabilities p_5 and p_6 more closely resembles the standard setting in which a self-generated anchor is established for the first assessment, which is subsequently adjusted for the second one. For these two probabilities, we found only a single pair of assessments in which the direction of adjustment was (presumably) incorrect.

6 Conclusions

As part of the engineering efforts for a Bayesian network for the early detection of Classical Swine Fever in pigs, we elicited a limited number of conditional probabilities from 38 pig experts and veterinary practitioners from six European countries outside the Netherlands. The main goal of the elicitation was to investigate whether our Bayesian network constructed with Dutch experts, appropriately reflected the practices and insights of veterinary experts across Europe. For our investigations, we obtained a total of 58 series of probability assessments, pertaining to two groups of related conditional probabilities. In this paper, we investigated numerical and qualitative properties of the assessments obtained. Of all analysed numerical properties, the modal interval appeared to be the most robust, although even for this property only limited consensus was found. The studied qualitative properties proved to be far more consistent among assessors and across countries, and matched the properties embedded in the original Dutch assessments upon which our Bayesian network builds. This finding suggests that at least the qualitative properties captured in our network have sufficient support in other European countries.

Our finding that the assessments per probability show considerable variance, yet embed consistent qualitative properties, may be attributed to an anchoring-and-adjustment heuristic applied by the veterinary experts who were not experienced probability assessors. They most likely based their first assessment on an estimate of how often they had experienced the presented scenario; their varying expertise caused large variation in these first assessments. After the initial setting of the anchor, their further reasoning was mostly likely based upon expertise, causing the subsequent assessments to go in the correct direction. Our findings suggest that, in general, when a series of probabilities have to be assessed, subjects had best first establish an ordering on related probabilities. For the actual assessment task, the probabilities then are best presented in the ordering agreed upon. If at all possible, the subjects should be provided with a reliable anchor for the first assessment. Variation in individual assessments from multiple experts nonetheless is bound to occur because of differences in background and experience.

References

[1] A.R.W. Elbers, A. Stegeman, H. Moser, H.M. Ekker, J.A. Smak, F.H. Pluimers. The classical swine fever epidemic 1997–1998 in the Netherlands: descriptive epidemiology. *Preventive Veterinary Medicine*, vol. 42, pp. 157 – 184, 1999.

[2] N. Epley. A tale of Tuned Decks? Anchoring as accessibility and anchoring as adjustment. In D. J. Koehler and N. Harvey, editors, *The Blackwell Handbook of Judgment and Decision Making*, Blackwell Publishers, Oxford, UK, pp. 240 – 256, 2004.

[3] L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B. Aleman and B.G. Taal, 'How to elicit many probabilities', in: K.B. Laskey and H. Prade (eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, San Francisco, California, pp. 647 – 654, 1999.

[4] K.E. Jacowitz and D. Kahneman. Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, vol. 21, pp. 1161 – 1166, 1995.

[5] P.N. Johnson-Laird, P. Legrenzi, V. Girotto, M.S. Legrenzi and J.-P. Caverni. Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, vol. 106, pp. 62 – 88, 1999.

[6] H. Levy. *Stochastic Dominance. Investment Decision Making under Uncertainty*. Studies in Risk and Uncertainty 12, Springer, New York, 2006.

[7] S. Renooij and C.L.M. Witteman. Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, vol. 22, pp. 169 – 194, 1999.