Experiences with Eliciting Probabilities from Multiple Experts

Linda C. van der Gaag[1], Silja Renooij[1], Hermi J.M. Schijf[1],
Armin R. Elbers[2], and Willie L. Loeffen[2]

[1] Department of Information and Computing Sciences, Utrecht University, NL
[2] Department of Virology, Central Veterinary Institute of Wageningen UR, NL

Abstract. Bayesian networks are typically designed in collaboration with a single domain 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 enforcing 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 domains governed by biological processes. While much attention has focused on algorithms for learning Bayesian networks from data, our experiences with 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, for both the engineers involved and 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 wider use.

In collaboration with two experts from the Central Veterinary Institute in the Netherlands, we are in the process of developing a decision-support system to supply veterinary practitioners with an independent 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. For its design, in-depth interviews were held with the two participating experts and case reviews were conducted with eight Dutch swine practitioners. 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.

S. Greco et al. (Eds.): IPMU 2012, Part III, CCIS 299, pp. 151–160, 2012.

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. Bayesian networks in fact are often intended for wider use than just by the experts with whom they are 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 pig experts and veterinary practitioners in six European countries outside the Netherlands. During these meetings, we were granted time with the experts to discuss some details of the current network. Among other information, we gathered assessments for a limited number of conditional probabilities for our network. Our intention was not to elicit 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 are supposed to hold universally across countries. We thus mimicked a realistic elicitation setting and 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 experts in six countries. We investigated the assessments obtained for the separate probabilities by establishing summary statistics, both per country and across countries. We further studied the series of assessments obtained and the qualitative properties of dominance embedded in them. We found large variances among the numerical assessments per probability, both within and between countries. Much higher consistency was found for the embedded dominance properties, however. Apparently, this type of qualitative information is more robust than numerical information. 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.

The present paper reports our findings and experiences from the project meetings. In Sect. 2, we briefly introduce the background of our application. Sect. 3 describes the set-up of the meetings and the elicitation method used. Sect. 4 summarises the assessments obtained. In Sect. 5, we analyse our findings from a qualitative perspective. The paper ends with our reflections in Sect. 6.

2 The Context

In 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 an infectious viral disease with a potential for rapid spread through contact between infected and non-infected susceptible pigs. When a pig is first infected, it will show an increased body temperature and a sense of malaise. 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

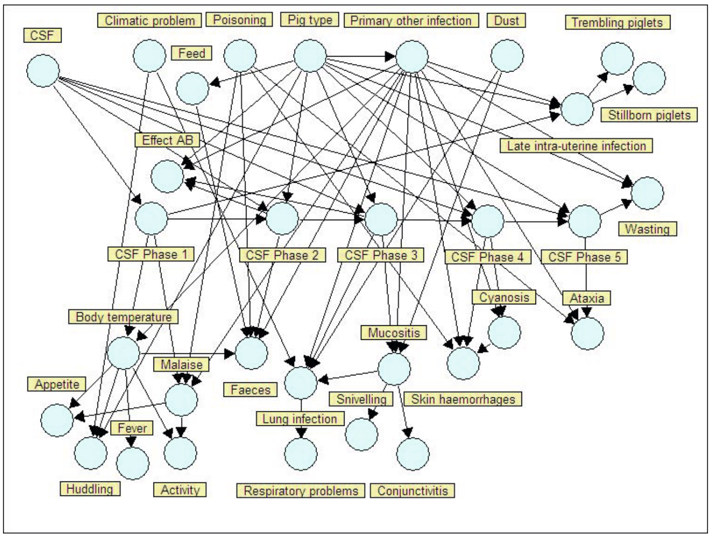

Fig. 1. The graphical structure of the Bayesian network for the early detection of CSF

ultimately cause the pig to die. 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 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 CSF project, we are developing a decision-support system to supply veterinary practitioners with an 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. Fig. 1 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 at renowned veterinary institutes in Belgium, Denmark, Germany, Great-Britain, Italy, and Poland. For each meeting, a small number of experts were invited from all over the host country; the invitees ranged from veterinary pig practitioners to researchers conducting experimental CSF infection studies. During the meetings, we were granted some time to discuss details of the CSF network. Within the allotted time, the experts were presented with a lecture about the working of the network; in addition, the assessment task to be performed was introduced.

Consider a pig *without an infection of the mucous* in the upper respiratory tract. How likely is it that this pig shows tear marks as a result of a *conjunctivitis* ?

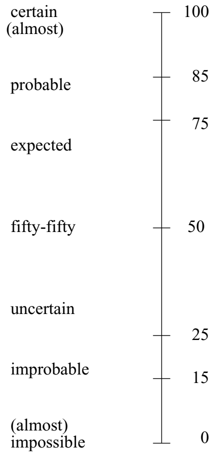

Fig. 2. A fragment of text for a requested probability, and the accompanying scale

For the assessment task, a tailored elicitation method was used, in which a requested probability was presented to the assessor as a fragment of text stated in veterinary terms and accompanied by a vertical scale with numerical and verbal anchors as illustrated in Fig. 2; for further details of the elicitation method, we refer to [1]. The assessor was asked to carefully consider the fragment of text and to indicate his assessment by marking the scale. The use of the probability scale was demonstrated during the plenary introduction of the task.

For our investigations, we selected twelve probabilities. In the present paper, we focus on the six probabilities summarised in Table 1; for the other six probabilities similar results were found. The six probabilities from the table were elicited from the experts in the displayed order. The probabilities p_1 through p_4 denote the probabilities of finding the 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 version of the 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 a mucous infection

Table 1. The six probabilities discussed in this paper, with the original assessments

Probability	*Original 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

in the upper respiratory tract, respectively; these two probabilities define 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 elicitations for the network's construction.

With the set-up outlined above, we obtained assessments for the probabilities p_1 through p_6 from a total of 38 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 Taking a Quantitative Perspective: Summary Statistics

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

4.1 The Data Obtained, the Analyses and the Results

Upon studying the responses obtained from our elicitation efforts, we found that the experts had used different methods for indicating their assessments on the probability scale. 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 assessments for further analysis. Some experts, however, had encircled one of the verbal anchors positioned beside the scale. Since the anchors indicate a fuzzy probability range [2], these circles were not used for numerical analysis. We obtained 58 complete 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 assessments obtained, which included 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

Table 2. Ranges, means $\overline{x}$ and standard deviations s of the assessments for the probability p_1 per country; assessments and means in bold lie in 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. Means $\overline{x}$ and standard deviations s of the assessments for the probabilities $p_2, \ldots, p_6$ per country; means in bold lie 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)

are provided in Table 3. We further computed some statistics per probability over all countries, which are summarised in Table 4. Note that the overall mean per probability may differ from the mean of the country means as a result of 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 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 $p_1, \ldots, p_4$ and p_6, we further performed post-hoc testing of pairwise equality under the assumption of equal variances. Post-hoc testing for p_6 did not reveal any significant pairwise differences of the means per country. For the probabilities $p_1, \ldots, p_4$, 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 differ 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. 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 to the atmosphere in the group. Among 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 within the focused area of Classical Swine Fever: it is well known from the theory of naive probability [3], that probability estimates are influenced by the assessor's experience. An interesting finding in this respect is that in some countries the assessments for the first probability p_1 were rather close to one another, while in other countries larger ranges were found; this closeness of assessments may point

Table 4. Ranges, modal intervals *mod* with frequencies # and means $\overline{x}$ with standard deviations s of all assessments per probability; means in bold lie in the 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)

to similarities in background and experience, yet may also be explained from a bias introduced by someone remarking out loud that some scenario, for example, is quite likely. Another explanation for the observed differences lies in the commonly used *anchoring-and-adjustment* heuristic: using this heuristic, people 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 [4,5]. 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-adjustment heuristic.

While the above observations can explain the variation among assessors, they do not explain the observed differences between countries. Remarkable differences were found, for example, for the means for each of the probabilities $p_1, \ldots, p_4$ from country $\mathcal{C}$, compared to the 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, “CSF doesn’t exist 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 we got strong indications from an independent native speaker who afterwards listened to the recording, that the translations were not always 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 modal intervals found in the other countries. A likely explanation for this finding may be that our expert provided his assessments from an entirely different background: 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 on the spot and obvious errors or inconsistencies could thus 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.

To conclude we would like to mention that without tailored experimentation in a more controlled elicitation setting, no definite conclusions can be drawn about the origins of the observed differences in the probability assessments.

5 Taking a Qualitative Perspective: Stochastic Dominance

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 address the qualitative properties embedded in the series of assessments.

5.1 The Data Obtained, the Analyses and the Results

For our qualitative analysis, we had available the same 58 series of numerical assessments from which we established standard statistics in the previous section. In addition to these numerical series, we had also available 10 complete sets of verbal assessments, that is, assessments composed of encircled verbal anchors from the probability scale. We observe that while we could not use these assessments in our quantitative analysis, the stability of the rank order of the verbal anchors does allow studying their qualitative properties [2].

For the qualitative analysis, we observe that although the six probabilities under study are probabilistically independent, they are not so from a domain point of view. 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. The 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 obtained.

For studying dominance properties, a total ordering of the conditioning contexts in the series of probabilities under study is required. For the probabilities $p_1, \ldots, p_4$ therefore a total ordering of the 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 'no-muco' $<$ 'muco' for the conditioning contexts. In addition, a total ordering of the probabilities themselves is required. For the numerical assessments, the standard numerical ordering is used. For the verbal assessments, we took the ordering on the verbal anchors from the probability scale, that is, we assumed '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 note that the assessments from our Dutch expert exhibit these properties.

For the probabilities $p_1, \ldots, p_4$, the assessments of 18 of the 29 experts (62%) who gave a complete numerical series, were found to obey the expected dominance property. In seven series, a 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 four remaining 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 $p_1, \ldots, p_4$, 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 obtained series of assessments are far more consistent among the individual experts and across countries, than the statistics studied in Sect. 4. For the probabilities $p_1, \ldots, p_4$ for example, a relatively large number of experts (62%) matched 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-adjustment 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 earlier assessment for p_1; for the final probability in the series, again an adjustment towards a higher value was performed, to beyond the assessment for p_2. Also of interest is the finding that six violations of the property of dominance among the probabilities $p_1, \ldots, p_4$ were caused not by an adjustment in the wrong direction but by a wrong amount. 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. For the probabilities p_5 and p_6, the direction of adjustment was (presumably) incorrect for a single pair of assessments.

6 Conclusions

As part of engineering 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 goal of the elicitation was to gain insight in the extent to which our Dutch expert-based network reflected the practices and insights of veterinary experts across Europe. All in all, we obtained 58 series of probability assessments, pertaining to two groups of related conditional probabilities. In this paper, we investigated summary statistics over the separate assessments and studied properties of stochastic dominance embedded in the assessment series. While the statistics showed only limited consensus, the dominance properties proved to be far more consistent among assessors and across countries. This finding suggests that at least the properties of stochastic dominance captured in our network have sufficient support in other European countries.

To our best knowledge, anchoring and adjusting has not been studied in tasks where a series of more than two related probabilities is assessed. It is unknown therefore, whether people would typically use the first anchor for all subsequent assessments, or tie each assessment to the previous one. Insights in the strategies which are commonly used by assessors can come only from carefully controlled experiments. Based upon our experiences and pending experimental evidence, we propose that assessors first establish a stable ordering on a series of related probabilities; the probabilities subsequently are presented in the ordering agreed upon. By thus prefixing the ordering of the probabilities, order violations ensuing from incorrect amounts of adjustment are forestalled. If at all possible, moreover, the assessors had best be provided with at least one reliable anchor, for example based upon literature or estimated from a rich enough data collection. Variation in individual assessments from multiple experts nonetheless is bound to occur because of differences in background and experience.

Acknowledgement. We would like to thank the Classical Swine Fever experts participating in the EPIZONE project meetings for their willingness to partake in the probability assessment task.

References

1. van der Gaag, L.C., Renooij, S., Witteman, C.L.M., Aleman, B., Taal, B.G.: How to elicit many probabilities. In: Laskey, K.B., Prade, H. (eds.) Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 647–654. Morgan Kaufmann Publishers, San Francisco (1999)
2. Renooij, S., Witteman, C.L.M.: Talking probabilities: communicating probabilistic information with words and numbers. International Journal of Approximate Reasoning 22, 169–194 (1999)
3. Johnson-Laird, P.N., Legrenzi, P., Girotto, V., Legrenzi, M.S., Caverni, J.-P.: Naive probability: A mental model theory of extensional reasoning. Psychological Review 106, 62–88 (1999)
4. Epley, N.: A tale of Tuned Decks? Anchoring as accessibility and anchoring as adjustment. In: Koehler, D.J., Harvey, N. (eds.) The Blackwell Handbook of Judgment and Decision Making, pp. 240–256. Blackwell Publisher, Oxford (2004)
5. Jacowitz, K.E., Kahneman, D.: Measures of anchoring in estimation tasks. Personality and Social Psychology Bulletin 21, 1161–1166 (1995)
6. Levy, H.: Stochastic Dominance. Investment Decision Making under Uncertainty. Studies in Risk and Uncertainty, vol. 12. Springer, New York (2006)