ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-489

Knowledge-based Bias Correction — A Case Study in Veterinary Decision Support

Thomas E. Krak[1] and **Linda C. van der Gaag**

Abstract. In collaboration with experts from veterinary research institutes throughout Europe, we developed a decision-support system for the early detection of Classical Swine Fever in pigs. For evaluating our system's diagnostic performance, practitioners and researchers collected data from the real-world field and from laboratory experiments. Originating from different sources, these data could not be viewed as constituting an unbiased sample from a single probability distribution. In this paper, we present a knowledge-based method for correcting the biases in estimates from such divergent data. We demonstrate the use of our method for estimating the sensitivity and specificity characteristics of our veterinary decision-support system.

1 INTRODUCTION

In close collaboration with veterinary experts from the research institutes involved in the European EPIZONE network of excellence, we developed an early-warning system for Classical Swine Fever (CSF) in individual pigs. Classical Swine Fever is a highly infectious viral disease, which is notifiable by law. Upon its detection, broad-scoped eradication measures are installed, with possibly major economic consequences. Our system is aimed at supplying veterinary practitioners with an independent tool for identifying suspect patterns of disease as early on in an outbreak of CSF as possible.

Embedded in our system is a Bayesian network for establishing the posterior probability of the clinical symptoms of an individual animal being caused by Classical Swine Fever. The performance of this network is studied in terms of its sensitivity and specificity characteristics, which describe the network's ability to distinguish between CSF-infected animals and diseased animals without CSF. These characteristics would ideally be determined from real-world data of both infected and non-infected animals. Since the European Union is currently free of Classical Swine Fever however, data from CSF-infected animals cannot be collected from the field setting in which the Bayesian network is to be employed. For establishing the network's sensitivity and specificity characteristics therefore, data were obtained from different sources. Data from animals without CSF were collected by pig veterinarians upon visiting pig farms with disease problems of unknown cause. Data pertaining to animals with a CSF infection were collected by veterinary researchers from inoculation experiments in a high-containment laboratory setting. All data were collected using the same standardised protocol.

Since our Bayesian network is to be employed in veterinary practice, its performance is investigated for real-world pig farms. The performance on diseased animals without CSF is readily established from the collected field data. An estimate of the network's performance on CSF-infected animals can in essence be obtained from the laboratory data submitted by the veterinary researchers. This latter estimate cannot be considered unbiased with respect to the real-world field setting, however. While animals with the disease present with the same CSF-specific pattern of clinical symptoms regardless of the setting, the field and laboratory settings differ considerably in for example the distribution of animal types and environment conditions.

Motivated by the above considerations for our domain of application, we address in this paper the problem of establishing unbiased probability estimates from datasets involving systematic bias. We show that by exploiting domain knowledge, unbiased distributions can effectively be obtained by weighting the available data with case-specific correction factors. We present a general method for this purpose and demonstrate its use for estimating the performance characteristics of our Bayesian network for Classical Swine Fever.

The paper is organised as follows. Section 2 provides some background information on our application domain and introduces the CSF network; in Section 3 we describe the collected data. Section 4 presents our method for establishing unbiased probability estimates from systematically biased data in general, and Section 5 details its application for estimating unbiased performance characteristics. Section 6 reports the sensitivity and specificity of our CSF network, as established by means of our method. The paper ends with our concluding observations and directions for further research in Section 7.

2 AN EARLY-WARNING SYSTEM FOR CSF

Classical Swine Fever is a viral pig disease with a potential for rapid spread. The early signs of the disease are quite aspecific, and are often attributed to an intestinal or respiratory infection. When the disease progresses however, it is associated with an accumulating failure of body systems, which will ultimately cause the animal to die. The disease is notifiable by law, which means that any suspicion of its presence has to be reported immediately to the agricultural authorities; control measures, involving closure of the farm, are then installed. The longer a CSF infection remains undetected, the longer the virus can circulate without hindrance, both within a herd and between herds. Because of the major economic consequences of an outbreak of the disease, reducing the high-risk period of time between first infection of a herd and first detection is of primary importance.

In collaboration with experts from the research institutes participating in the EPIZONE network of excellence, we developed a Bayesian network for the early detection of Classical Swine Fever in pigs. For its construction, we held in-depth interviews with the veterinary experts; in addition, case reviews were conducted with swine practitioners, both with and without clinical CSF experience. The

[1] Department of Information and Computing Sciences, Utrecht University, The Netherlands; email: {T.E.Krak, L.C.vanderGaag}@uu.nl

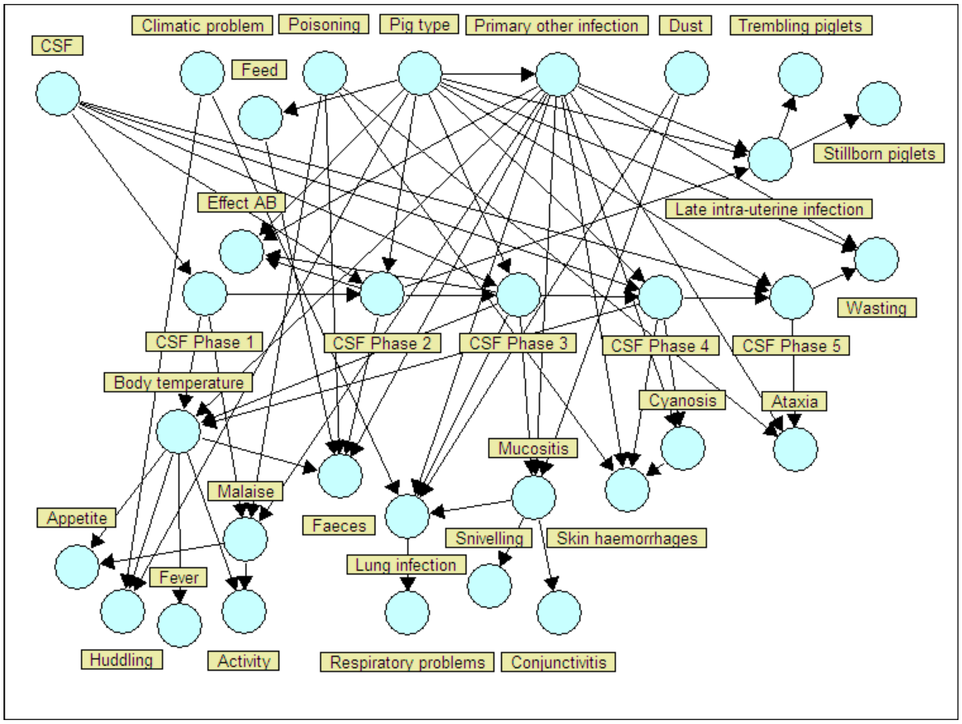

Figure 1. The graphical structure of the Bayesian network for the early detection of Classical Swine Fever in individual pigs.

graphical structure of the network, which includes 32 random variables, is shown in Figure 1. About half of the network's variables describe clinical symptoms which are relevant for either confirming or ruling out a diagnosis of Classical Swine Fever; another six variables serve to organise these symptom variables into important combinations pertaining to different phases in the presentation of the disease [2]. The remaining variables describe the internal effects of a CSF infection and alternative explanations for observed symptoms. The dependencies among the variables are described by 67 arcs, which are quantified by some 1300 (conditional) probabilities.

The CSF network is aligned to veterinarians visiting pig farms. The network thus takes clinical evidence only for its input, and does not require pathology findings or results from laboratory tests. It focuses on individual animals and takes for its input the symptoms found in a specific pig; it further takes type information about the animal and some information about pen conditions. Based on the entered evidence, the network establishes the posterior probability of the symptoms of the animal being caused by an infection with the CSF virus. We note that the network does not take information about the clinical pattern exhibited by a specific pig over time for its input, as individual pigs are not readily identifiable within a herd.

In the present paper, we address establishing the sensitivity and specificity of our CSF network. These commonly-used characteristics convey information about the performance quality of a diagnostic system in general: where the sensitivity of a system equals the probability of diagnosing an individual as suffering from a disease when it actually does have the disease, the system's specificity is defined as the probability of diagnosing an individual without the disease as indeed not having the disease. To study these performance characteristics for our CSF network, the concept of diagnosis needs to be formalised in terms of calculated posterior probabilities. We compare to this end a posterior probability computed from the network, against a pre-set threshold probability. If the probability of CSF for a specific pig exceeds this threshold probability, we say that the animal is diagnosed as suffering from Classical Swine Fever. We would like to note that the established performance characteristics cannot be expected to convey high quality of our Bayesian network, as it is not intended for stand-alone use: the CSF network is embedded in a more involved model, which takes the pattern and rate of spreading of clinical symptoms throughout a herd into consideration in addition to information from selected individual animals.

3 DATA COLLECTION

For evaluating the diagnostic performance of our Bayesian network for Classical Swine Fever, we collected a range of real-world data.

In a two-year field study in the Netherlands, eleven veterinarians were asked to collect information from up to five individual pigs upon visiting a herd with disease problems of unknown cause. They were asked more specifically to gather data on 15 clinical symptoms per animal; for this purpose, the practitioners were supplied with a personal digital assistent running a standardised protocol [3]. During the study, data were collected from 375 pigs. Veterinarians from the partner countries of the EPIZONE network were also invited to collect and submit field data; these practitioners were supplied with a paper version of the data-entry screens of the pda used by the Dutch veterinarians. The EPIZONE partners submitted data from yet another 45 animals. All in all therefore, data from a total of 420 pigs were collected. We would like to note that, since the European Union is currently free of Classical Swine Fever, all collected data came from animals without the disease and can thus be used for establishing an estimate of the specificity of our network only.

To evaluate the sensitivity of our Bayesian network for Classical Swine Fever, researchers from the national veterinary laboratories in

Table 1. The distribution of pig types in the collected data.

	Field data	*Laboratory data*
Suckling piglet	40	10
Weaned piglet	106	64
Finishing pig	229	26
Sow	36	2
Boar	9	–
Total	420	102

the EPIZONE countries were asked to collect data from their CSF experiments. In such an experiment, one or more animals from among a close-contact group of pigs are inoculated with a specific CSF strain, after which all individuals are monitored over time; the goal of the experiment typically is to gain evidence of the rate of infection and of the progression of disease. The researchers were asked to record data from each animal in their experiment, according to the protocol used in the field trial; in line with the goals of the experiment, data were recorded every two or three days. Over a period of three years, information was collected from 23 inoculation experiments, involving a total of 128 animals. The information revealed that 26 of these animals did not show any clinical symptoms on any of the recording days, even though they had been in close contact with CSF-infected animals. Since our Bayesian network takes clinical information only and will be used by veterinarians upon encountering disease problems, we decided to remove the data from these individuals, leaving us with data from 102 pigs for evaluating our network's sensitivity.

The data available for studying the performance characteristics of our Bayesian network thus originate from two different sources which cannot be viewed as embedding the same probability distribution over all random variables concerned. Specifically, even though CSF-infected animals present with the same clinical pattern regardless of the setting, the field and laboratory settings differ in their distribution of animal types and environment conditions. While the real-world pig husbandry includes all animal types, ranging from suckling piglets to boars, the individuals used in inoculation experiments are of less divergent type; this difference in type distribution is reflected in the data, as illustrated in Table 1. Also, in the laboratory setting, environment conditions are much more controlled than in the field setting. As a consequence, the data from the two different information sources cannot be simply combined into a single dataset from which both the sensitivity and the specificity of our Bayesian network for Classical Swine Fever can be estimated.

4 KNOWLEDGE-BASED BIAS CORRECTION

Motivated by the considerations of systematic bias for our application, we developed a general method for estimating unbiased probability distributions from datasets involving known bias.

4.1 Debiasing probability distributions

We consider a (multi-)set $\mathcal{D}$ of cases which are described by discrete random variables. We distinguish an outcome variable of interest Y; for ease of exposition, we assume this variable to be binary, and write y and $\bar{y}$ to indicate positive and negative cases respectively. The set of random variables describing the relevant features of the cases will be denoted by $\mathbf{X}$; we will use $\Omega_{\mathbf{X}}$ to denote the set of possible value combinations for $\mathbf{X}$. We assume that $\mathbf{X}$ is partitioned into a set $\mathbf{X}_s$ of symptom variables and a set $\mathbf{X}_t$ of type variables, with the associated sets of value combinations $\Omega_{\mathbf{X}_s}, \Omega_{\mathbf{X}_t}$ respectively, with $\Omega_{\mathbf{X}} = \Omega_{\mathbf{X}_s} \times \Omega_{\mathbf{X}_t}$; for our application for example, the variables from $\mathbf{X}_s$ describe animal-specific clinical evidence, while $\mathbf{X}_t$ captures animal type, feed quality and environment conditions.

Over the variables $\mathbf{X}, Y$, we assume two probability distributions, one of which describes the occurrence of cases in the field and the other one pertains to the laboratory setting. We introduce a new binary random variable L to distinguish between the two distributions; the value l is used to indicate the laboratory setting and $\bar{l}$ indicates the field. In essence, we are interested in the probability distribution $\Pr(\mathbf{X}, Y \mid \bar{l}\,)$, that is, in the distribution over the variables $\mathbf{X}$, Y as it exists in the real-world field setting. The dataset $\mathcal{D}$ available for estimating the distribution of interest includes cases from both the field and the laboratory setting. We note that only if the probability distributions in the field and in the laboratory are the same, can estimates for the distribution of interest be obtained directly from this dataset. For our application we know however, that the distributions in the field and in the laboratory are *not* the same. Any estimates obtained from the dataset $\mathcal{D}$ thus need to be corrected for the differences between the two distributions. We introduce a binary random variable S to accommodate for the systematic bias in the available data; this variable indicates whether or not a particular observation over $\mathbf{X}, Y, L$ could in principle be included in the dataset.

We now address the problem of estimating the probability distribution in the field from a dataset $\mathcal{D}$ which includes both field data and laboratory data as described above. More specifically, we present a general method for estimating from such a dataset the conditional distributions $\Pr(\mathbf{X} \mid Y, \bar{l}\,)$ over the feature variables for negative cases and positive cases in the field, repectively. Our method is tailored to applications in which

- $\Pr(\mathbf{X}, Y \,|\, l\,) \neq \Pr(\mathbf{X}, Y \,|\, \bar{l}\,)$, that is, the probability distributions in the field and in the laboratory setting differ;
- $\Pr(s \mid y, \bar{l}\,) = \Pr(s \mid \bar{y}, l\,) = 0$, that is, we cannot observe any positive cases in the field nor any negative cases in the laboratory.

We note that the dataset $\mathcal{D}$ allows direct estimation of the conditional probability distribution $\Pr(\mathbf{X} \mid \bar{y}, \bar{l}\,)$ over the feature variables for negative field cases; since all negative cases included in $\mathcal{D}$ are known to have originated from the field, these cases were drawn directly from the probability distribution of interest. The dataset does not provide for estimating the conditional probability distribution $\Pr(\mathbf{X} \mid y, \bar{l}\,)$ over the feature variables for positive field cases. Since all positive cases are known to have come from laboratories, just the probability distribution $\Pr(\mathbf{X} \mid y, l)$ can be estimated directly from $\mathcal{D}$. Under mild conditions however, can the systematic bias in the latter distribution be corrected, to thereby provide an approximation of the yet unknown probability distribution $\Pr(\mathbf{X} \mid y, \bar{l}\,)$ over the feature variables for positive cases in the field.

We are interested in the conditional probability distribution $\Pr(\mathbf{X} \mid y, \bar{l}\,)$ over the feature variables, for which we have that

$$\Pr(\mathbf{X} \mid y, \bar{l}\,) \quad = \quad \Pr(\mathbf{X}_s \mid \mathbf{X}_t, y, \bar{l}\,) \cdot \Pr(\mathbf{X}_t \mid y, \bar{l}\,)$$

We address the two terms in the right-hand side of the expression separately, and focus first on the term $\Pr(\mathbf{X}_s \mid \mathbf{X}_t, y, \bar{l}\,)$ which captures the probability distribution over the symptom variables in positive field cases, per case type. We assume that the symptoms observed in positive laboratory cases are representative for positive cases that would be found in the field, for any case type; we would like to note that this assumption is a realistic one to make for our application as it underlies the very goal of performing laboratory experiments to study patterns of animal disease. By this assumption, we find that

$$\Pr(\mathbf{X}_s \mid \mathbf{X}_t, y, \bar{l}\,) \quad = \quad \Pr(\mathbf{X}_s \mid \mathbf{X}_t, y, l\,)$$

We further assume that the selection of cases for inclusion in the dataset $\mathcal{D}$ is *not* dependent of the symptoms observed. Building upon this assumption, we find that

$$\Pr(\mathbf{X}_s \mid \mathbf{X}_t, y, l\,) \;=\; \Pr(\mathbf{X}_s \mid \mathbf{X}_t, s, y, l\,)$$

The probability distribution $\Pr(\mathbf{X}_s \mid \mathbf{X}_t, s, y, l\,)$ thus arrived at describes the distribution over the symptom variables, per case type, for positive cases collected from the laboratory. We note that this probability distribution is readily estimated from the dataset $\mathcal{D}$.

We now turn to the second term in the expression for the distribution $\Pr(\mathbf{X} \mid y, \bar{l}\,)$ of interest, that it, we address the probability distribution $\Pr(\mathbf{X}_t \mid y, \bar{l}\,)$ over the type variables in positive field cases. In general, we have that

$$\begin{aligned}\Pr(\mathbf{X}_t \mid \bar{l}\,) \;=\; & \Pr(\mathbf{X}_t \mid y, \bar{l}\,) \cdot \Pr(y \mid \bar{l}\,) + \\ & \Pr(\mathbf{X}_t \mid \bar{y}, \bar{l}\,) \cdot \Pr(\bar{y} \mid \bar{l}\,)\end{aligned}$$

Assuming that the true probability distribution over the outcome variable in the field is strictly positive, we find that

$$\Pr(\mathbf{X}_t \mid y, \bar{l}\,) \;=\; \frac{\Pr(\mathbf{X}_t \mid \bar{l}\,) - \Pr(\mathbf{X}_t \mid \bar{y}, \bar{l}\,) \cdot \Pr(\bar{y} \mid \bar{l}\,)}{\Pr(y \mid \bar{l}\,)}$$

We would like to note that this assumption again is quite realistic for our application, since early warning pertains to the detection of actually possible diseases. The probability distribution $\Pr(\mathbf{X}_t \mid \bar{y}, \bar{l}\,)$ in the expression above is readily established from the available negative field data. The distribution $\Pr(Y \mid \bar{l}\,)$ over the outcome variable in the field however, cannot be estimated from the data. For this probability distribution, we resort to domain knowledge and assume that an estimate of the prior probability of finding a positive case in the field can be obtained, either from the scientific literature or from experts. A similar assumption is made for the probability distribution $\Pr(\mathbf{X}_t \mid \bar{l}\,)$ over the type variables in the field.

Building upon the above considerations, we conclude that the probability distribution of interest is estimated as

$$\begin{aligned}\Pr(\mathbf{X} \mid y, \bar{l}\,) \;&=\; \Pr(\mathbf{X} \mid s, y, l) \cdot \frac{\Pr(\mathbf{X}_t \mid y, \bar{l}\,)}{\Pr(\mathbf{X}_t \mid s, y, l)} \\ &=\; \Pr(\mathbf{X} \mid s, y, l) \cdot \left(\frac{\Pr(\mathbf{X}_t \mid \bar{l}\,) - \Pr(\mathbf{X}_t \mid \bar{y}, \bar{l}\,) \cdot \Pr(\bar{y} \mid \bar{l}\,)}{\Pr(\mathbf{X}_t \mid s, y, l) \cdot \Pr(y \mid \bar{l}\,)} \right)\end{aligned}$$

under the following assumptions:

- the selection of cases for inclusion in the dataset $\mathcal{D}$ is *not* biased in $\mathbf{X}_s$, given any type information, outcome status and setting, that is, $(S \perp\!\!\!\perp \mathbf{X}_s \mid \mathbf{X}_t, Y, L)$;
- the symptoms observed in positive laboratory cases are representative for positive cases that would be observed in the field, given the cases' type information, that is, $(L \perp\!\!\!\perp \mathbf{X}_s \mid \mathbf{X}_t, y)$;
- the true distribution $\Pr(Y \mid \bar{l}\,)$ is strictly positive.

We would like to note that in the derivation above, we also built on the assumption that the distribution $\Pr(\mathbf{X}_t \mid s, y, l)$ of observed types in positive laboratory cases is strictly positive. If this assumption does not hold, we know beforehand that the estimates obtained for the probability distribution $\Pr(\mathbf{X} \mid y, \bar{l}\,)$ will not constitute good approximations. We will return to this observation in Section 5.

4.2 Related work

The problem of bias correction is studied widely. The general question focused on is how to correct probability estimates for a bias that was introduced through a data-collection regime by which a case's selection is not independent of its features and/or outcome. Since researchers are often confronted with such a selection bias in practice, a large corpus of literature has been published in which this question is addressed under various assumptions and for different applications; for examples we refer to [4, 5, 6, 7]. The approach taken by most researchers is to estimate a model of the selection probability based on the feature variables in which the data are biased. This model is then used to compute weights for the contribution of individual cases to unbiased estimates of a quantity of interest, such as the parameters of a regression model. The various methods proposed differ in how the selection model is estimated and how the weights are computed, as well as in the applications to which they are tailored.

As the application specifics and underlying assumptions of most methods for dealing with sample selection bias are quite different from ours, we cannot directly apply them to the problem addressed in the present paper. While available methods establish a scalar selection probability to compute the weights for individual data cases, our method requires a probability distribution over the variables $\mathbf{X}_t$ in which the data are biased. We recall moreover, that the data are biased not just in the type variables, but in the outcome variable Y as well. More specifically, the model describing the selection bias in our data is $\Pr(s \mid \mathbf{X}_t, y, \bar{l}\,)$. Since in our application the selection probability equals zero for all case types, no informative weights can be computed from the selection probabilities as is assumed by available methods. For our method therefore, we resorted to assuming further independences to allow the computation of weights from the distribution over the type variables $\mathbf{X}_t$ instead.

5 FINDING UNBIASED CHARACTERISTICS

Our method for knowledge-based bias correction described above can be used for any computations for which the probability distributions $\Pr(\mathbf{X}, Y \mid \bar{l}\,)$ need to be available. In this section, we demonstrate, as an example of its application, how the method is used for establishing unbiased performance characteristics for a diagnostic system. We recall that the performance of such a system is generally expressed by its sensitivity and specificity. A system's specificity is defined as the proportion of true negative cases which the system singles out as indeed being negative; its sensitivity is the proportion of true positive cases which the system identifies as being positive.

The performance characteristics of a diagnostic system are typically estimated from a dataset of positive and negative cases originating from a single probability distribution. To this end, the system is looked upon as implementing a function $\hat{y}$ which establishes for each case $\mathbf{x}$ over the feature variables $\mathbf{X}$ a value prediction $\hat{y}(\mathbf{x})$ for the outcome variable Y. The sensitivity of the system is then expressed more formally as $\mathbb{E}_{\mathbf{x}|y}\left[\iota^{+}(\hat{y}(\mathbf{x}))\right]$, where the indicator function ι^{+} is defined as $\iota^{+}(\hat{y}(\mathbf{x})) = 1$ if $\hat{y}(\mathbf{x}) = y$, and $\iota^{+}(\hat{y}(\mathbf{x})) = 0$ otherwise; the system's specificity is expressed similarly, through an indicator function ι^{-}. From an unbiased dataset $\mathcal{D}$ of positive and negative cases, the sensitivity of the system would be estimated as

$$\widehat{\mathbb{E}}_{\mathbf{x}|y}\left[\iota^{+}(\hat{y}(\mathbf{x}))\right] \;=\; \frac{1}{|\mathcal{D}_y|} \cdot \sum_{\mathbf{x} \in \mathcal{D}_y} \iota^{+}(\hat{y}(\mathbf{x}))$$

where $\mathcal{D}_y$ is the subset of positive cases from $\mathcal{D}$ and where individual occurrences of cases in $\mathcal{D}_y$ are counted separately. A similar expression is obtained for the system's specificity.

We now suppose that for estimating the performance characteristics of a specific diagnostic system, we have available not an unbiased dataset, but a dataset $\mathcal{D}$ involving systematic bias as described in the previous section. From this dataset, we readily establish an unbiased estimate of the system's specificity for the field as

$$\widehat{\mathbb{E}}_{\mathbf{x}\mid\bar{y},\bar{l}}\left[\iota^-(\hat{y}(\mathbf{x}))\right] \;=\; \frac{1}{|\mathcal{D}_{\bar{y},\bar{l}}|}\cdot\sum_{\mathbf{x}\in\mathcal{D}_{\bar{y},\bar{l}}}\iota^-(\hat{y}(\mathbf{x}))$$

where $\mathcal{D}_{\bar{y},\bar{l}}$ is the subset of negative field cases from $\mathcal{D}$. To obtain an estimate of the system's ability to correctly identify positive field cases, we need to correct the distribution $\Pr(\mathbf{X} \mid s, y, l)$ estimated from the dataset $\mathcal{D}$, for the laboratory bias. With the property

$$\frac{\Pr(\mathbf{X}\mid y,\bar{l})}{\Pr(\mathbf{X}_t\mid y,\bar{l})} = \frac{\Pr(\mathbf{X}\mid s,y,l)}{\Pr(\mathbf{X}_t\mid s,y,l)}$$

derived in Section 4.1, we have for the system's field sensitivity that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\mid y,\bar{l}}\left[\iota^+(\hat{y}(\mathbf{x}))\right] \;&=\; \sum_{\mathbf{x}\in\Omega_{\mathbf{X}}}\iota^+(\hat{y}(\mathbf{x}))\cdot\Pr(\mathbf{x}\mid y,\bar{l})\\
&= \sum_{\mathbf{x}_t\in\Omega_{\mathbf{X}_t}}\Pr(\mathbf{x}_t\mid y,\bar{l})\cdot\sum_{\mathbf{x}_s\in\Omega_{\mathbf{X}_s}}\iota^+(\hat{y}(\mathbf{x}))\cdot\frac{\Pr(\mathbf{x}\mid y,\bar{l})}{\Pr(\mathbf{x}_t\mid y,\bar{l})}\\
&= \sum_{\mathbf{x}_t\in\Omega_{\mathbf{X}_t}}\Pr(\mathbf{x}_t\mid y,\bar{l})\cdot\sum_{\mathbf{x}_s\in\Omega_{\mathbf{X}_s}}\iota^+(\hat{y}(\mathbf{x}))\cdot\frac{\Pr(\mathbf{x}\mid s,y,l)}{\Pr(\mathbf{x}_t\mid s,y,l)}\\
&= \sum_{\mathbf{x}_t\in\Omega_{\mathbf{X}_t}}\Pr(\mathbf{x}_t\mid y,\bar{l})\cdot\mathbb{E}_{\mathbf{x}_s\mid\mathbf{x}_t,s,y,l}\left[\iota^+(\hat{y}(\mathbf{x}))\right]
\end{aligned}$$

where $\mathbf{x}$ is taken consistent with $\mathbf{x}_s$, $\mathbf{x}_t$. The field sensitivity is now estimated from the data through

$$\begin{aligned}
\widehat{\mathbb{E}}_{\mathbf{x}\mid y,\bar{l}}\left[\iota^+(\hat{y}(\mathbf{x}))\right] &= \sum_{\mathbf{x}_t\in\Omega_{\mathbf{X}_t}}\widehat{\Pr}(\mathbf{x}_t\mid y,\bar{l})\cdot\widehat{\mathbb{E}}_{\mathbf{x}_s\mid\mathbf{x}_t,s,y,l}\left[\iota^+(\hat{y}(\mathbf{x}))\right]\\
&= \sum_{\mathbf{x}_t\in\Omega_{\mathbf{X}_t}}\widehat{\Pr}(\mathbf{x}_t\mid y,\bar{l})\cdot\sum_{\mathbf{x}_s\in\mathcal{D}_{\mathbf{x}_t,y,l}}\frac{\iota^+(\hat{y}(\mathbf{x}))}{|\mathcal{D}_{\mathbf{x}_t,y,l}|}\\
&= \frac{1}{|\mathcal{D}_{y,l}|}\cdot\sum_{\mathbf{x}_t\in\Omega_{\mathbf{X}_t}}\widehat{\Pr}(\mathbf{x}_t\mid y,\bar{l})\cdot\sum_{\mathbf{x}_s\in\mathcal{D}_{\mathbf{x}_t,y,l}}\frac{\iota^+(\hat{y}(\mathbf{x}))}{\widehat{\Pr}(\mathbf{x}_t\mid s,y,l)}\\
&= \frac{1}{|\mathcal{D}_{y,l}|}\cdot\sum_{\mathbf{x}\in\mathcal{D}_{y,l}}\iota^+(\hat{y}(\mathbf{x}))\cdot\frac{\widehat{\Pr}(\mathbf{x}_t\mid y,\bar{l})}{\widehat{\Pr}(\mathbf{x}_t\mid s,y,l)}
\end{aligned}$$

We note that the estimates $\widehat{\Pr}(\mathbf{x}_t \mid s, y, l)$ are readily obtained from the available data. Domain knowledge further provides the estimates $\widehat{\Pr}(\mathbf{x}_t \mid y, \bar{l})$, as described in Section 4.1. From the above derivation, we conclude that debiasing the sensitivity estimate obtained from the dataset thus amounts to weighting the contribution of each case by the case-specific factor $\widehat{\Pr}(\mathbf{x}_t \mid y, \bar{l}) \,/\, \widehat{\Pr}(\mathbf{x}_t \mid s, y, l)$.

In Section 4.1 we already mentioned that if the observed distribution $\Pr(\mathbf{X}_t \mid s, y, l)$ over the types involved in positive laboratory cases is not strictly positive, we know that the estimates obtained for the type distribution $\Pr(\mathbf{X}_t \mid y, \bar{l})$ in positive field cases will not be good approximations. The decomposition of the sensitivity estimate in terms conditional on the type variables $\mathbf{X}_t$ shows that this property holds unabatedly for the unbiased sensitivity as well: if very few cases of a particular type have been recorded, then the term for the associated conditional will not be reliable. We further note that if particular case types are missing altogether from the dataset, then the sensitivity estimate obtained can never reach the value 1, not even if we would have $\hat{y}(\mathbf{x}) = y$ for all $\mathbf{x} \in \Omega_{\mathbf{X}}$. In view of missing case types therefore, our debiasing method yields a lower bound on a system's sensitivity. Knowledge of the distribution $\Pr(\mathbf{X}_t \mid y, \bar{l})$ then provides also for establishing an upper bound on the sensitivity. This upper bound is computed by taking the lower bound as described above and adding the proportion of unobserved case types as they are known to occur in the field; we note that the thus established upper bound reflects the assumption that these cases would all be classified correctly. Knowledge of the distribution can further be used to compute a point estimate of the unbiased sensitivity by assuming that the sensitivity estimate for the observed case types is representative for the entire field; the point estimate is computed by dividing the established lower bound by the proportion of observed case types as they occur in the field. We note that this point estimate serves to normalize perfect classification on the data to yield a sensitivity estimate equal to 1. We would like to emphasize that while these approaches correct for missing case types, they do not serve to correct for types with a small yet non-zero number of cases. For such types, it may be worthwhile to widen the established bounds by removing the associated cases from the data, rather than letting their unreliable contributions influence the estimate obtained for the system's sensitivity.

6 APPLICATION TO THE CSF NETWORK

To establish unbiased performance characteristics for our Bayesian network for Classical Swine Fever, we applied our method for knowledge-based bias correction to the collected pig data. The laboratory data were pre-processed for this purpose. We removed for each animal the recordings of all days on which it revealed no or just a single clinical symptom. This pre-processing step was motivated by our early-warning system being aimed at use on farms with disease problems: an attending veterinarian would not use the system for animals showing hardly any clinical symptoms. Because the data collected from the inoculation experiments include multiple recordings per pig pertaining to different days moreover, we performed uniform random sub-sampling to remove the dependencies between these recordings. Furthermore, since the laboratory data included information from two sows only, also these recordings were removed from the dataset, as suggested in Section 5. For each pig case from the resulting dataset, the posterior probability of the clinical symptoms being caused by a CSF infection was computed from the network and subsequently compared against a threshold value α as described in Section 2; if and only if the probability computed for a specific animal exceeded the threshold value α, was the animal taken as being diagnosed with Classical Swine Fever. In view of the very small prior probability of the occurrence of CSF in the field, we used quite small threshold values α in our evaluation study.

Before our method of bias correction could be applied, estimates for a number of probabilities had to be available. We recall from Section 4.1, that the method requires the probability distribution $\Pr(\mathbf{X}_t \mid \bar{l})$ over the type variables in the field and the prior probability $\Pr(y \mid \bar{l})$ of finding Classical Swine Fever in the real-world setting. Since these required probabilities had already been obtained from domain experts upon quantifying the CSF network, they were readily available for our current purposes. We further had to establish the type variables in which the laboratory data were biased. Based upon knowledge of the field and laboratory settings, we concluded that these data were biased in the animal type, the presence of climatic problems, and the composition of the animals' feed.

In our evaluation study of the sensitivity and specificity characteristics of the CSF network, sub-sampling and performance estimation were repeated by a 100 runs. Figure 2 plots the performance characteristics of the CSF network for different threshold probabilities α. The reported specificity was computed from the collected field data.

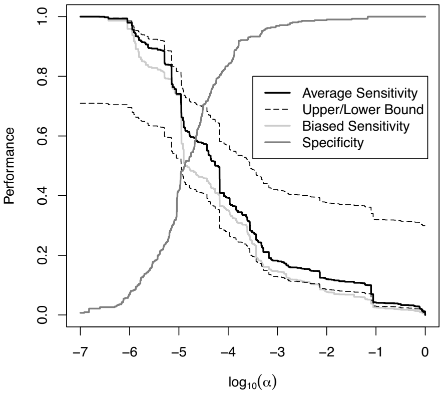

Figure 2. Unbiased estimates of the sensitivity of the CSF network, expressed as the average point estimate and upper/lower bounds, for various threshold values; the specificity and biased sensitivity are also shown.

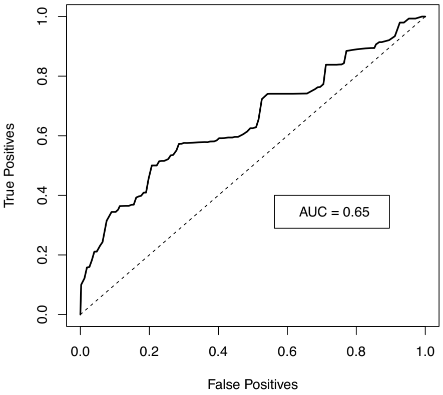

Figure 3. The ROC curve of the CSF network based on its specificity and unbiased sensitivity point estimate; the area under the curve (AUC) is 0.65.

The figure further reports the average unbiased sensitivity over the range of threshold values; in addition, upper and lower bounds on the sensitivity are shown, to accommodate for the absence of sows, boars, and climatic and feed problems from the laboratory population. The figure also plots the biased sensitivity calculated from the data. We note that the bias from the laboratory setting shows a tendency to underestimate the network's detection abilities. For completeness, the unbiased performance characteristics are also reported numerically in Table 2, again for various threshold probabilities; the table further reports the standard deviation of the (unbiased) lower bound, established from the repeated sub-sampling of the laboratory data. To conclude, Figure 3 summarizes the overall performance of the CSF network by depicting the ROC curve computed from the network's specificity and unbiased sensitivity point estimate.

Table 2. Numerical values of the specificity and unbiased sensitivity estimates for various probability thresholds.

α	SPEC	SENS_{LOW}	SENS_{AVG}	$\text{SENS}_{\text{HIGH}}$	$\sigma_{\text{SENS}_{\text{LOW}}}$
0.00001	0.42	0.53	0.74	0.82	0.029
0.00005	0.77	0.37	0.52	0.66	0.032
0.0001	0.84	0.28	0.39	0.57	0.033
0.0005	0.95	0.16	0.23	0.45	0.028
0.001	0.97	0.13	0.18	0.42	0.023
0.005	0.99	0.11	0.15	0.40	0.022
0.01	0.99	0.09	0.12	0.38	0.022
0.05	0.99	0.08	0.11	0.37	0.022
0.1	1.00	0.03	0.04	0.32	0.019

7 CONCLUSIONS AND FUTURE RESEARCH

Motivated by the difficulty of establishing reliable estimates for the performance characteristics of our real-world Bayesian network, we studied the problem of correcting probability distributions estimated from an available dataset for known systematic biases. We presented a general method which, under mild conditions, serves to effectively debias estimated probability distributions by exploiting domain knowledge. In essence, our method amounts to establishing case-specific correction factors to be used for weighting case contributions to a quantity of interest.

Although our method has broader applicability than just for establishing the performance characteristics of our Bayesian network, it is tailored to a specific type of application. Our method assumes for example that the positive and negative cases to be distinguished originate from strictly separated settings. While for many problems in real-world application domains the assumption of a zero-inclusion probability will be satisfied, the scope of practicability of our method would be broadened if it were able to deal with settings in which the inclusion probabilities are indegenerate. Our future research efforts will be directed to enhancing our debiasing method to this end.

Acknowledgements. We are most grateful to Armin Elbers and Willie Loeffen from the Central Veterinary Institute, Wageningen UR, the Netherlands, for their invaluable help with the construction of our Bayesian network for Classical Swine Fever. We would further like to thank the veterinary practitioners and researchers who put considerable effort in collecting data from individual pigs for the purpose of evaluating the performance of our network.

REFERENCES

[1] A. Elbers, L.C. van der Gaag, S. Schmeiser, A. Uttenthal, L. Lohse, J. Nielsen, H. Crooke, S. Blome, W.L. Loeffen. A Bayesian clinical decision support system for early detection of Classical Swine Fever in individual pigs — Evaluation of the sensitivity and specificity of the model. In: *The Eighth International Pestivirus Symposium of the European Society for Veterinary Virology*, Hannover: 104, 2011.

[2] L.C. van der Gaag, J. Bolt, W.L. Loeffen, A. Elbers. Modelling patterns of evidence in Bayesian networks: a case-study in Classical Swine Fever. In: E. Hüllermeier, R. Kruse, F. Hoffmann (editors). *Computational Intelligence for Knowledge-based Systems Design*, LNAI vol. 6178, Springer-Verlag, Berlin: 675-684, 2010.

[3] L.C. van der Gaag, H.J.M. Schijf, A.R. Elbers, W.L. Loeffen. Preserving precision as a guideline for interface design for mathematical models. In: J.W.H.M. Uiterwijk, N. Roos, M.H.M. Winands (editors). *Proceedings of the 24th Benelux Conference on Artificial Intelligence*, Maastricht University: 107–114, 2012.

[4] J.J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, **47**: 153–161, 1979.

[5] P.R. Rosenbaum, D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**: 41–55, 1983.

[6] C. Winship, R.D. Mare. Models for sample selection bias. *Annual Review of Sociology*, **18**: 327–350, 1992.

[7] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In: *Proceedings of the Twenty-first International Conference on Machine Learning*: 114, 2004.