The Impact of Overconfidence Bias on Practical Accuracy of Bayesian Network Models: An Empirical Study

Marek J. Drużdżel[1,2] **& Agnieszka Oniśko**[1,3]
[1] Faculty of Computer Science, Białystok Technical University, Wiejska 45A, 15-351 Białystok, Poland
[2] Decision Systems Laboratory, School of Information Sciences and Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA
[3] Magee Womens Hospital, University of Pittsburgh Medical Center, Pittsburgh, PA 15260, USA

Abstract

In this paper, we examine the influence of overconfidence in parameter specification on the performance of a Bayesian network model in the context of Hepar II, a sizeable Bayesian network model for diagnosis of liver disorders. We enter noise in the parameters in such a way that the resulting distributions become biased toward extreme probabilities. We believe that this offers a systematic way of modeling expert overconfidence in probability estimates. It appears that the diagnostic accuracy of Hepar II is less sensitive to overconfidence in probabilities than it is to underconfidence and to random noise, especially when noise is very large.

1 INTRODUCTION

Decision-analytic methods provide an orderly and coherent framework for modeling and solving decision problems in decision support systems [5]. A popular modeling tool for complex uncertain domains is a Bayesian network [13], an acyclic directed graph quantified by numerical parameters and modeling the structure of a domain and the joint probability distribution over its variables. There exist algorithms for reasoning in Bayesian networks that typically compute the posterior probability distribution over some variables of interest given a set of observations. As these algorithms are mathematically correct, the ultimate quality of reasoning depends directly on the quality of the underlying models and their parameters. These parameters are rarely precise, as they are often based on subjective estimates. Even when they are based on data, they may not be directly applicable to the decision model at hand and be fully trustworthy.

Search for those parameters whose values are critical for the overall quality of decisions is known as sensitivity analysis. Sensitivity analysis studies how much a model output changes as various model parameters vary through the range of their plausible values. It allows to get insight into the nature of the problem and its formalization, helps in refining the model so that it is simple and elegant (containing only those factors that matter), and checks the need for precision in refining the numbers [8]. It is theoretically possible that small variations in a numerical parameter cause large variations in the posterior probability of interest. Van der Gaag and Renooij [17] found that practical networks may indeed contain such parameters. Because practical networks are often constructed with only rough estimates of probabilities, a question of practical importance is whether overall imprecision in network parameters is important. If not, the effort that goes into polishing network parameters might not be justified, unless it focuses on their small subset that is shown to be critical.

There is a popular belief, supported by some anecdotal evidence, that Bayesian network models are overall quite tolerant to imprecision in their numerical parameters. Pradhan et al. [14] tested this on a large medical diagnostic model, the CPCS network [7, 16]. Their key experiment focused on systematic introduction of noise in the original parameters (assumed to be the gold standard) and measuring the influence of the magnitude of this noise on the average posterior probability of the true diagnosis. They observed that this average was fairly insensitive to even very large noise. This experiment, while ingenious and thought provoking, had two weaknesses. The first of these, pointed out by Coupé and van der Gaag [3], is that the experiment focused on the average posterior rather than individual posterior in each diagnostic case and how it varies with noise, which is of most interest. The second weakness is that the posterior of the correct diagnosis is by itself not a sufficient measure of model robustness. The weaknesses of this experiment were also discussed in [6] and [9]. In our earlier work [9], we replicated the experiment of Pradhan et al. using

HEPAR II, a sizeable Bayesian network model for diagnosis of liver disorders. We systematically introduced noise in HEPAR II's probabilities and tested the diagnostic accuracy of the resulting model. Similarly to Pradhan et al., we assumed that the original set of parameters and the model's performance are ideal. Noise in the original parameters led to deterioration in performance. The main result of our analysis was that noise in numerical parameters started taking its toll almost from the very beginning and not, as suggested by Pradhan et al., only when it was very large. The region of tolerance to noise, while noticeable, was rather small. That study suggested that Bayesian networks may be more sensitive to the quality of their numerical parameters than popularly believed. Another study that we conducted more recently [4] focused on the influence of progressive rounding of probabilities on model accuracy. Here also, rounding had an effect on the performance of HEPAR II, although the main source of performance loss were zero probabilities. When zeros introduced by rounding are replaced by very small non-zero values, imprecision resulting from rounding has minimal impact on HEPAR II's performance.

Empirical studies conducted so far that focused on the impact of noise in probabilities on Bayesian network results disagree in their conclusions. Also, the noise introduced in parameters was usually assumed to be random, which may not be a reasonable assumption. Human experts, for example, often tend to be overconfident [8]. This paper describes a follow-up study that probes the issue of sensitivity of model accuracy to noise in probabilities further. We examine whether a bias in the noise that is introduced into the network makes a difference. We enter noise in the parameters in such a way that the resulting distributions become biased toward extreme probabilities. We believe that this offers a systematic way of modeling expert overconfidence in probability estimates. Our results show again that the diagnostic accuracy of HEPAR II is sensitive to imprecision in probabilities. It appears, however, that the diagnostic accuracy of HEPAR II is less sensitive to overconfidence in probabilities than it is to random noise. We also test the sensitivity of HEPAR II to underconfidence in parameters and show that underconfidence in paramaters leads to more error than random noise.

The remainder of this paper is structured as follows. Section 2 introduces the HEPAR II model. Section 3 describes how we introduced noise into our probabilities. Section 4 describes the results of our experiments. Finally, Section 5 discusses our results in light of previous work.

2 THE HEPAR II MODEL

Our experiments are based on HEPAR II [10, 11], a Bayesian network model consisting of over 70 variables modeling the problem of diagnosis of liver disorders. The model covers 11 different liver diseases and 61 medical findings, such as patient self-reported data, signs, symptoms, and laboratory tests results. The structure of the model, (i.e., the nodes of the graph along with arcs among them) was built based on medical literature and conversations with domain experts and it consists of 121 arcs. HEPAR II is a real model and it consists of nodes that are a mixture of propositional, graded, and general variables. There are on the average 1.73 parents per node and 2.24 states per variable. The numerical parameters of the model (there are 2,139 of these in the most recent version), i.e., the prior and conditional probability distributions, were learned from a database of 699 real patient cases. Readers interested in the HEPAR II model can download it from Decision Systems Laboratory's model repository at `http://genie.sis.pitt.edu/`.

As our experiments study the influence of precision of HEPAR II's numerical parameters on its accuracy, we owe the reader an explanation of the metric that we used to test the latter. We focused on diagnostic accuracy, which we defined in our earlier publications as the percentage of correct diagnoses on real patient cases. When testing the diagnostic accuracy of HEPAR II, we were interested in both (1) whether the most probable diagnosis indicated by the model is indeed the correct diagnosis, and (2) whether the set of w most probable diagnoses contains the correct diagnosis for small values of w (we chose a "window" of w=1, 2, 3, and 4). The latter focus is of interest in diagnostic settings, where a decision support system only suggest possible diagnoses to a physician. The physician, who is the ultimate decision maker, may want to see several alternative diagnoses before focusing on one.

With diagnostic accuracy defined as above, the most recent version of the HEPAR II model reached the diagnostic accuracy of 57%, 69%, 75%, and 79% for window sizes of 1, 2, 3, and 4 respectively [12].

3 INTRODUCTION OF NOISE INTO HEPAR II PARAMETERS

When introducing noise into parameters, we used essentially the same approach as Pradhan et al. [14], which is transforming each original probability into log-odds function, adding noise parametrized by a parameter σ (as we will show, even though σ is proportional to the amount of noise, in our case it cannot be directly interpreted as standard deviation), and trans-

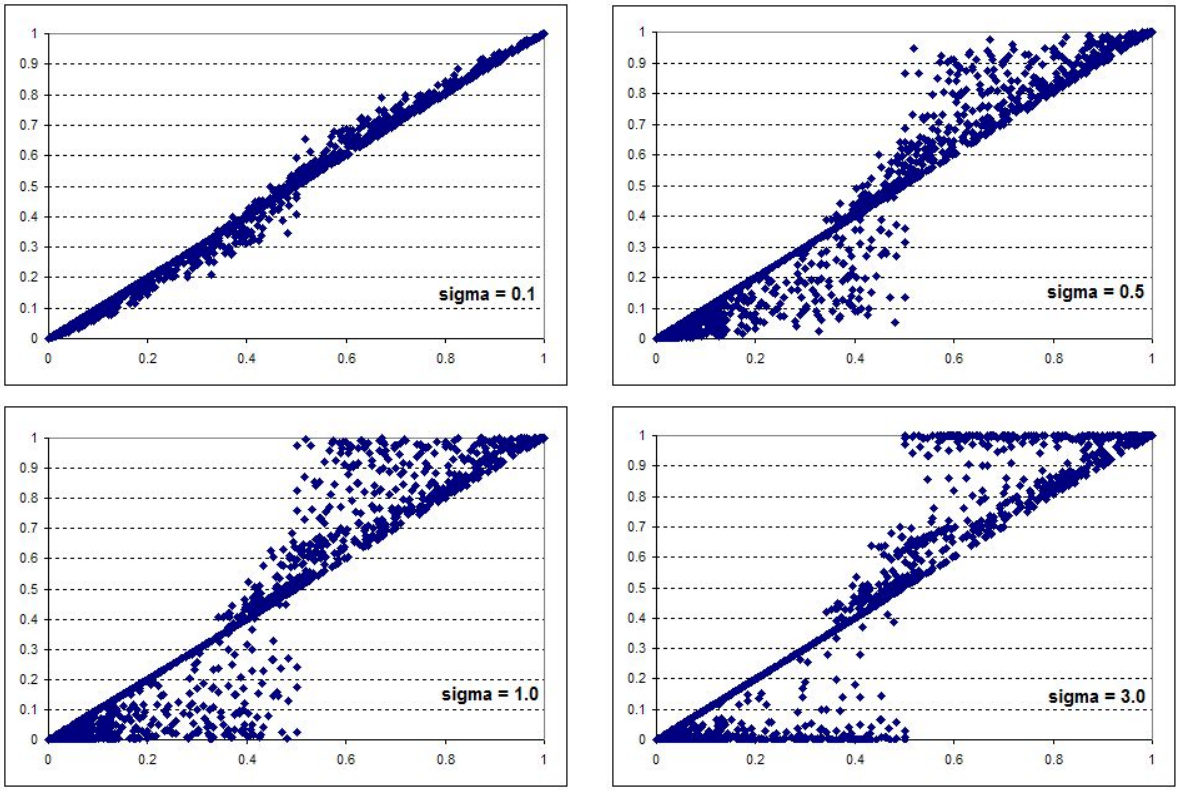

Figure 1: Transformed (biased, overconfident) vs. original probabilities for various levels of σ.

forming it back to probability, i.e.,

$$p' = Lo^{-1}[Lo(p) + \text{Noise}(0, \sigma)] \ , \tag{1}$$

where

$$Lo(p) = \log_{10}[p/(1-p)] \ . \tag{2}$$

3.1 Overconfidence bias

Now, we designed the Noise() function as follows. Given a discrete probability distribution Pr, we identify the smallest probability p_S. We transform this smallest probability p_S into p'_S by making it even smaller, according to the following formula:

$$p'_S = Lo^{-1}[Lo(p_S) - |\text{Normal}(0, \sigma)|] \ .$$

We make the largest probability in the probability distribution Pr, p_L larger by precisely the amount by which we decreased p_S, i.e.,

$$p'_L = p_L + p_S - p'_S \ .$$

We are by this guaranteed that the transformed parameters of the probability distribution Pr$'$ add up to 1.0.

Figure 1 shows the effect of introducing the noise. As we can see, the transformation is such that small probabilities are likely to become smaller and large probabilities are likely to become larger. Please note that distributions have become more biased towards the extreme probabilities. It is straightforward to prove that the entropy of Pr$'$ is smaller than the entropy of Pr. The transformed probability distributions reflect overconfidence bias, common among human experts.

An alternative way of introducing biased noise, suggested by one of the reviewers, is by means of building a logistic regression/IRT model (e.g., [1, 2, 15])for each conditional probability table and, subsequently, manipulating the slope parameter.

3.2 Underconfidence bias

Now, we designed the Noise() function as follows. Given a discrete probability distribution Pr, we identify the highest probability p_S. We transform this largest probability p_L into p'_L by making it smaller, according to the following formula:

$$p'_L = Lo^{-1}[Lo(p_L) - |\text{Normal}(0, \sigma)|] \ .$$

We make the smallest probability in the probability distribution Pr, p_S larger by precisely the amount by which we decreased p_L, i.e.,

$$p'_S = p_S + p_L - p'_L \ .$$

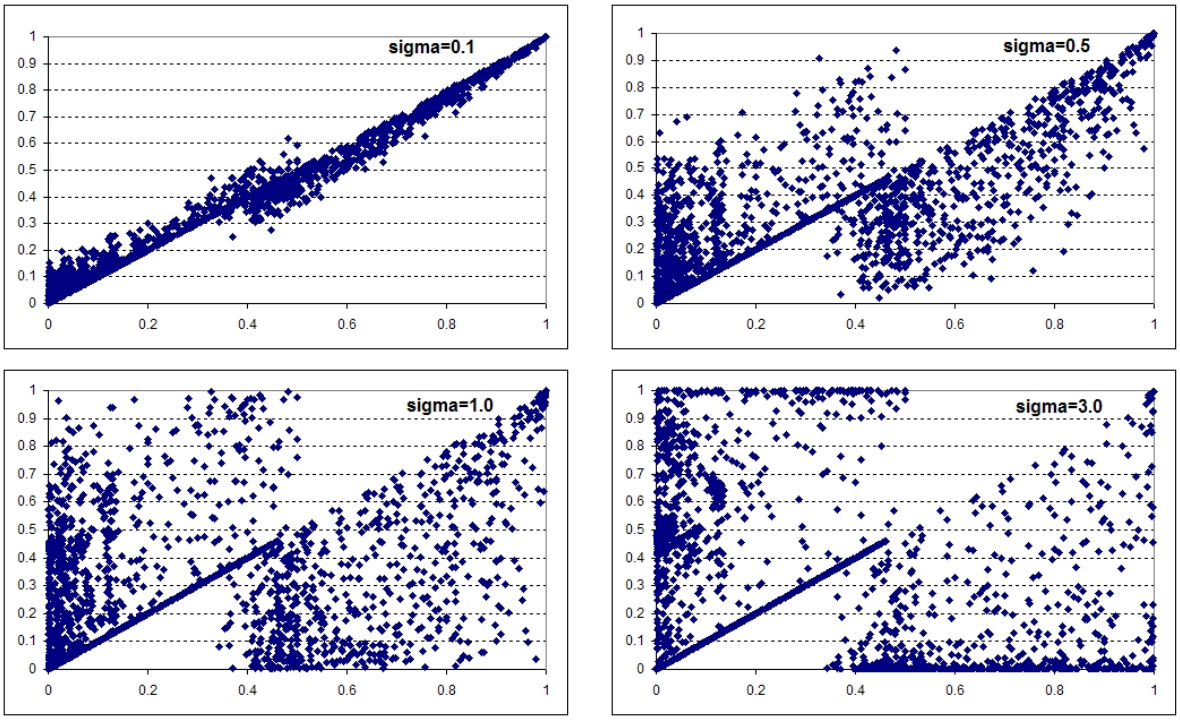

Figure 2: Transformed (biased, underconfident) vs. original probabilities for various levels of σ.

We are by this guaranteed that the transformed parameters of the probability distribution $\Pr'$ add up to 1.0.

Figure 2 shows the effect of introducing this noise. The transformed probability distributions reflect underconfidence bias.

3.3 Random noise

For illustration purpose, Figure 3 shows the transformation applied in our previous study [9]. For $\sigma > 1$ the amount of noise becomes so large that any value of probability can be transformed in any other value. This suggests strongly that $\sigma > 1$ is not really a region that is of interest in practice. The main reason why we look at such high σ values is that this was the range used in Pradhan et al. paper.

4 EXPERIMENTAL RESULTS

We have performed an experiment investigating the influence of biased noise in HEPAR II's probabilities on its diagnostic performance. For the purpose of our experiment, we assumed that the model parameters were perfectly accurate and, effectively, the diagnostic performance achieved was the best possible. Of course, in reality the parameters of the model may not be accurate and the performance of the model can be improved upon. In the experiment, we studied how this baseline performance degrades under the condition of noise, as described in Section 3.

We tested 30 versions of the network (each for a different standard deviation of the noise $\sigma \in < 0.0, 3.0 >$ with 0.1 increments) on all records of the HEPAR data set and computed HEPAR II's diagnostic accuracy. We plotted this accuracy in Figures 4 and 5 as a function of σ for different values of window size w.

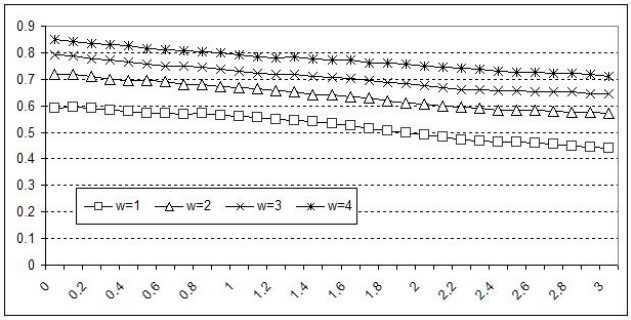

Figure 4: The diagnostic accuracy of HEPAR II for various window sizes as a function of the amount of biased overconfident noise (expressed by σ)
.

It is clear that HEPAR II's diagnostic performance deteriorates with noise. In order to facilitate comparison between biased and unbiased noise and, by this,

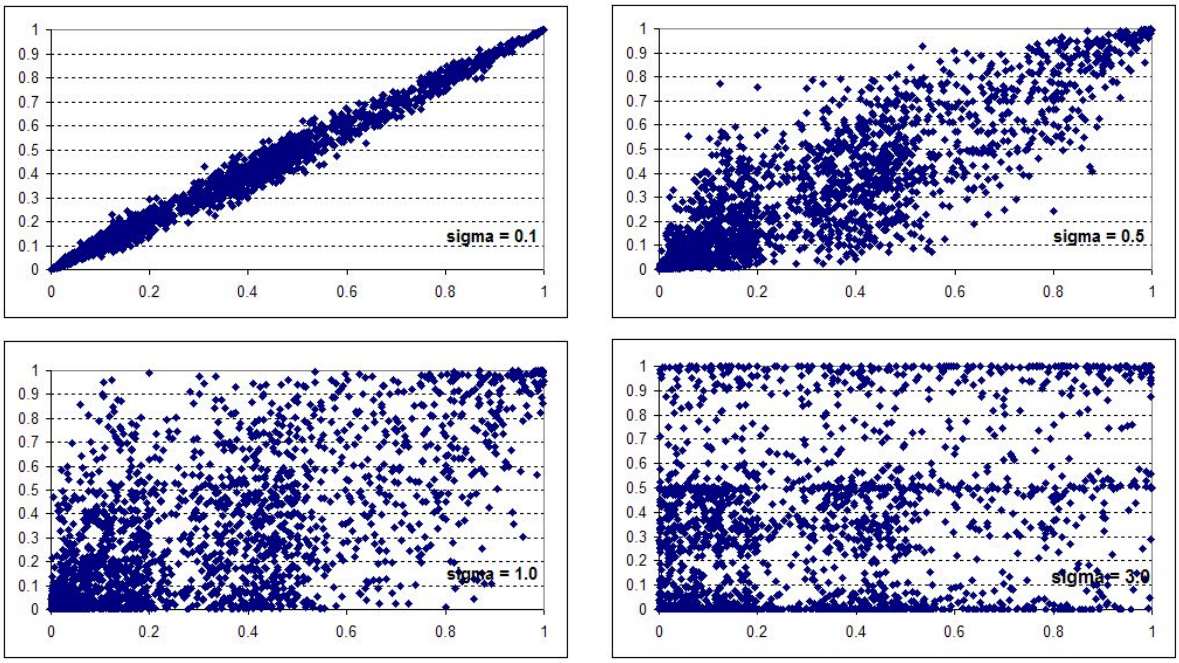

Figure 3: Transformed (unbiased) vs. original probabilities for various levels of σ.

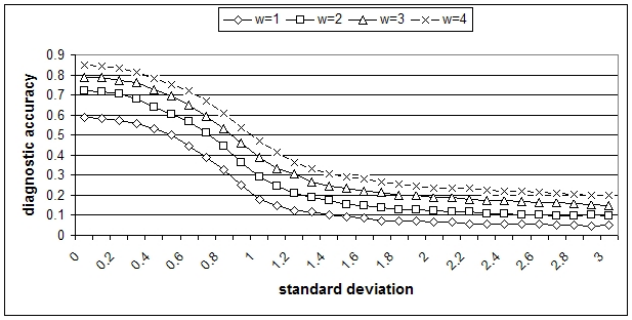

Figure 5: The diagnostic accuracy of HEPAR II for various window sizes as a function of the amount of biased underconfident noise (expressed by σ)
.

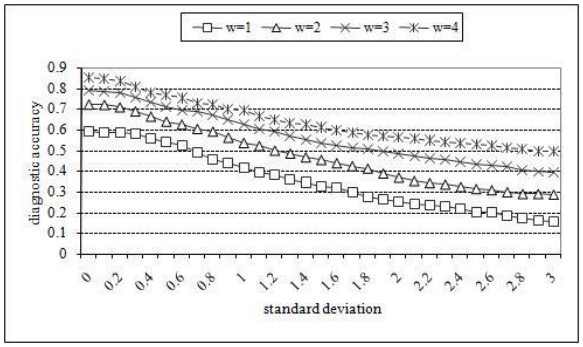

Figure 6: The diagnostic accuracy of HEPAR II for various window sizes as a function of amount of unbiased noise (expressed by σ) [9].

judgment of the influence of overconfidence bias on the results, we reproduce the experimental result of [9] in Figure 6. The results are qualitatively similar, although it can be seen that performance under overconfidence bias degrades more slowly with the amount of noise than performance under random noise. Performance under underconfidence bias degrades the fastest of the three. Figure 7 shows the accuracy of HEPAR II ($w = 1$) for biased and unbiased noise on the same plot, where this effect is easier to see.

It is interesting to note that for small values of σ, such as $\sigma < 0.2$, there is only a minimal effect of noise on performance. This observation may offer some justification to the belief that Bayesian networks are not too sensitive to imprecision of their probability parameters.

5 SUMMARY

This paper has studied the influence of bias in parameters on model performance in the context of a practical medical diagnostic model, HEPAR II. We believe that the study was realistic in the sense of focusing on a real, context-dependent performance measure. Our study has shown that the performance of HEPAR II is sensitive to noise in numerical parameters, i.e., the diagnostic accuracy of the model decreases after introducing noise into numerical parameters of the model.

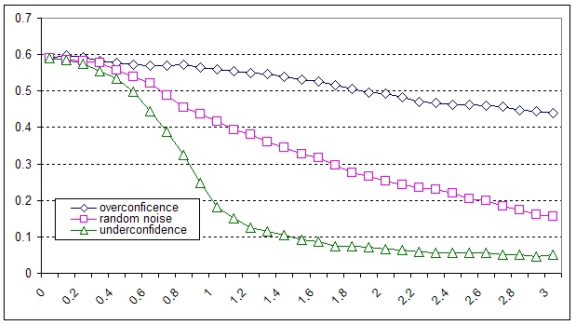

Figure 7: The diagnostic accuracy of HEPAR II as a function of the amount of noise (random, underconfident, and overconfident), window $w = 1$

While our result is merely a single data point that sheds light on the hypothesis in question, it looks like overconfidence bias has a smaller negative effect on model performance than random noise. Underconfidence bias leads to most serious deterioration of performance. While it is only a wild speculation that begs for further investigation, one might see our results as an explanation of the fact that humans tend to be overconfident rather than underconfident in their probability estimates.

Acknowledgments

This work was supported by the Air Force Office of Scientific Research grant FA9550-06-1-0243, by Intel Research, and by the MNiI (Ministerstwo Nauki i Informatyzacji) grant 3T10C03529. We thank Linda van der Gaag for suggesting extending our earlier work on sensitivity of Bayesian networks to precision of their numerical parameters by introducing bias in the noise. Reviewers for The Sixth Bayesian Modelling Applications Workshop provided several useful suggestions that have improved the readability and extended the scope of the paper.

The HEPAR II model was created and tested using SMILE, an inference engine, and GeNIe, a development environment for reasoning in graphical probabilistic models, both developed at the Decision Systems Laboratory, University of Pittsburgh, and available at `http://genie.sis.pitt.edu/`. We used SMILE in our experiments and the data pre-processing module of GeNIe for plotting scatter plot graphs in Figure 1.

References

[1] Russell G. Almond, Louis V. Dibello, F. Jenkins, R.J. Mislevy, D. Senturk, L.S. Steinberg, and D. Yan. Models for conditional probability tables in educational assessment. In T. Jaakkola and T. Richardson, editors, *Artificial Intelligence and Statistics 2001*, pages 137–143. Morgan Kaufmann, 2001.

[2] Russell G. Almond, Louis V. DiBello, Brad Moulder, and Juan-Diego Zapata-Rivera. Modeling diagnostic assessment with Bayesian networks. *Journal of Educational Measurement*, 44(4):341–359, 2007.

[3] Veerle H. M. Coupé and Linda C. van der Gaag. Properties of sensitivity analysis of Bayesian belief networks. *Annals of Mathematics and Artificial Intelligence*, 36:323–356, 2002.

[4] Marek J. Druzdzel and Agnieszka Oniśko. Are Bayesian networks sensitive to precision of their parameters? In S.T. Wierzchoń M. Kłopotek, M. Michalewicz, editor, *Intelligent Information Systems XVI, Proceedings of the International IIS'08 Conference*, pages 35–44, Warsaw, Poland, 2008. Academic Publishing House EXIT.

[5] Max Henrion, John S. Breese, and Eric J. Horvitz. Decision Analysis and Expert Systems. *AI Magazine*, 12(4):64–91, Winter 1991.

[6] O. Kipersztok and H. Wang. Another look at sensitivity analisis of Bayesian networks to imprecise probabilities. In *Proceedings of the Eight International Workshop on Artificial Intelligence and Statistics (AISTAT-2001)*, pages 226–232, San Francisco, CA, 2001. Morgan Kaufmann Publishers.

[7] B. Middleton, M.A. Shwe, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann, and G.F. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST–1/QMR knowledge base: II. Evaluation of diagnostic performance. *Methods of Information in Medicine*, 30(4):256–267, 1991.

[8] M. Granger Morgan and Max Henrion. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge, 1990.

[9] Agnieszka Oniśko and Marek J. Druzdzel. Effect of imprecision in probabilities on Bayesian network models: An empirical study. In *Working notes of the European Conference on Artificial Intelligence in Medicine (AIME-03): Qualitative and Model-based Reasoning in Biomedicine*, Protaras, Cyprus, October 18–22 2003.

[10] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. Extension of the Hepar II model to

multiple-disorder diagnosis. In S.T. Wierzchoń M. Kłopotek, M. Michalewicz, editor, *Intelligent Information Systems,* Advances in Soft Computing *Series*, pages 303–313, Heidelberg, 2000. Physica-Verlag.

[11] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. Learning Bayesian network parameters from small data sets: Application of Noisy-OR gates. *International Journal of Approximate Reasoning*, 27(2):165–182, 2001.

[12] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. An experimental comparison of methods for handling incomplete data in learning parameters of Bayesian networks. In S.T. Wierzchoń M. Kłopotek, M. Michalewicz, editor, *Intelligent Information Systems,* Advances in Soft Computing *Series*, Heidelberg, 2002. Physica-Verlag (A Springer-Verlag Company). 351–360.

[13] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1988.

[14] Malcolm Pradhan, Max Henrion, Gregory Provan, Brendan del Favero, and Kurt Huang. The sensitivity of belief networks to imprecise probabilities: An experimental investigation. *Artificial Intelligence*, 85(1–2):363–397, August 1996.

[15] Frank Rijmen. Bayesian networks with a logistic regression model for the conditional probabilities. *International Journal of Approximate Reasoning*, in press.

[16] M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann, and G.F. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST–1/QMR knowledge base: I. The probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30(4):241–255, 1991.

[17] Linda C. van der Gaag and Silja Renooij. Analysing sensitivity data from probabilistic networks. In *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2001)*, pages 530–537, San Francisco, CA, 2001. Morgan Kaufmann Publishers.