Application of decision-tree technique to assess herd specific risk factors for coliform mastitis in sows

Imke Gerjets,[1] Imke Traulsen,[1] Kerstin Reiners,[2] Nicole Kemper[3]

[1]Institute for Animal Breeding and Husbandry, Christian-Albrechts-University Kiel; [2]PIC Germany GmbH, Schleswig; [3]Institute of Agricultural and Nutritional Sciences, Martin-Luther-University Halle-Wittenberg, Halle, Germany

Abstract

This study aimed at investigating factors associated with coliform mastitis (CM) in sows, determined at herd level, by applying the decision-tree technique. Coliform mastitis represents an economically important disease in sows after farrowing that also affects the health, welfare and performance of the piglets. The decision-tree technique, a data mining method, may be an effective tool for making large datasets accessible and different sow herd information comparable. It is based on the C4.5-algorithm, which generates trees in a top-down recursive strategy. The technique can be used to detect weak points in farm management. Datasets of two farms in Germany, consisting of individual sow data, were analyzed and compared by decision-tree algorithms. Data were collected over the period April 2007-August 2010 from 987 sows of different parities (1-9) (499 sows with coliform mastitis and 488 healthy sows) and 596 sows (322 sows with coliform mastitis and 274 healthy sows), respectively. Depending on the dataset, different graphical trees were built showing relevant factors at the herd level, which may lead to coliform mastitis. The application of birth intervention and a higher number of piglets born alive and stillborn ones were the main risk factors identified by the decision-tree technique to be associated with coliform mastitis. Herd specific risk factors for the disease were illustrated what could prove beneficial in disease and herd management. The application of decision trees may be a possibility of analysing critical points and decisions in management on an individual farm basis.

Introduction

Coliform mastitis (CM) is an important infection in sows after farrowing followed by serious economic losses due to lower productivities of the affected sows and higher preweaning piglet mortalities.[1] The diseased animals suffer from fever and an inflammation of the mammary glands that often leads to a decreased milk secretion 24 to 48 h post partum. The average prevalence in herds is about 13%, but a prevalence up to 60% of the animals has also been reported.[2-4] The term coliform mastitis refers to a clinical mastitis due to coliform bacteria (*Escherichia species* (spp.), *Klebsiella* spp., *Enterobacter* spp. and *Citrobacter* spp.) which have been found to be associated with the disease complex in many studies.[1,3,5,6] As a multifactorial disease, CM is influenced by the strongly related main issues of management, feeding and hygiene as well as individual sow-related parameters.[7] It is generally assumed that optimal herd management including the detection of weak points is a key element in reducing the prevalence of diseases in general and of CM as multifactorial infection in herds in particular.[8] With the aid of management information technology, farmers are able to collect, process and interpret data based at individual animal level.[9] Data mining methods are special statistical instruments which are applied to detect relationships between attributes in datasets. The decision-tree technique, a data mining method, has been proven as an effective tool to make large farm datasets accessible and different sow herds information comparable.[10] The aim of this study was to investigate the application of the decision-tree technique to assess potential risk factors associated with CM-infected sows. Decision-trees which allow deduction of association rules could support the comparison and assessment of herd data and thereby the establishment of optimal and individual management strategies.

Correspondence: Imke Gerjets, Institute for Animal Breeding and Husbandry, Christian-Albrechts-Universität Kiel, Olshausenstraße 40, D-24098 Kiel, Germany.
Tel. +49.431.8807432 - Fax: +49.431.8805265.
E-mail: igerjets@tierzucht.uni-kiel.de

Key words: decision-tree modeling, management tool, mastitis, sows.

Acknowledgements: this research project was funded by the German Federal Ministry of Education and Research (BMBF) in the research programme FUGATO – Functional Genome Analysis in Animal Organisms, project geMMA – structural and functional analysis of the genetic variation of the MMA-syndrome (FKZ0315138).

Contributions: NK, study design and coordination; IG manuscript drafting, the statistical analysis and application of the decision-tree technique performing; IT statistical analysis participating; KR sow herds data providing. All authors read and approved the final manuscript.

Conflict of interest: the authors report no conflicts of interest.

Received for publication: 18 April 2011.
Accepted for publication: 6 May 2011.

This work is licensed under a Creative Commons Attribution 3.0 License (by-nc 3.0).

©Copyright I. Gerjets et al., 2010
Licensee PAGEPress, Italy
Veterinary Science Development 2011; 1:e6
doi:10.4081/vsd.2011.e6

Materials and Methods

Datasets

The study was based on datasets from two rearing herds in Germany with 1,200 (Farm A) and 1,800 sows (Farm B) collected from April 2008 to August 2010 within the scope of a microbiological study.[11] The farms were of high health status and tested free from porcine reproductive and respiratory syndrome-virus, rhinitis, *Actinobacillus pleuropneumoniae* dysentery and enzootic pneumonia.

The datasets comprised individual reproduction traits of the sows (Table 1) and a respective binary record of the occurrence of CM (present or absent). All sows were examined after farrowing and considered as mastitis cases when their rectal temperature was above the threshold of 39.5°C at 24 h *post-partum*[12] and the mammary glands showed definite signs of inflammation. Healthy half- or full-sib sows from the same farrowing group served as controls. The half-sib design was chosen due to further studies on the genetic background of CM via genotyping (Preissler *et al.*, *unpublished data*). Manual obstetric measures after the beginning of birth were defined as the trait

Table 1. Means (standard deviations) and frequencies (yes/no) of reproductive traits for Farms A and B.

Variable (abbreviation)	Farm A (n = 987)	Farm B (n = 596)
Number of parities per sow (*np*)	4.0 (1.9)	3.2 (1.9)
Piglets born alive per litter (*pba*)	12.1 (3.0)	12.3 (3.1)
Piglets born dead per litter (*pbd*)	1.2 (1.6)	1.0 (1.5)
Birth intervention (*biv*)	212/ 775	143/ 453
Birth induction (*bid*)	409/ 578	356/ 240

Non-commercial use only

birth intervention. Birth induction was the hormonal induction of birth after the 115. day of gestation in order to get the birth process started. The first dataset (Farm A) consisted of a total of 987 observations (animals) – 499 observations from CM-positive sows and 488 observations from CM-negative sows. The second dataset (Farm B) contained 596 observations whereas 322 observations distinguished CM-positive sows and 274 observations CM-negative sows. The mean number of parities per sow was 4.0 for Dataset A and 3.2 for Dataset B (Table 1). The average number of piglets born alive was 12.1 and 12.3 for Dataset A and B and the average number of stillborn piglets was 1.2 and 1.0 for Dataset A and B, respectively. The mean number of weaned piglets was 10.6 for both datasets.

Decision-tree algorithm

The C4.5-algorithm of the open source software WEKA was used to generate decision-trees by employing the top-down and recursive-splitting technique.[13] Every decision-tree consisted of a root node and internal nodes representing the attributes, and branches that characterized the attribute values. In this study, the reproduction parameters and the information of birth intervention (*biv*) and birth induction (*bid*) served as attributes. The leaves (leaf node of the decision-tree) expressed the binary decision (presence or absence of CM) and indicated the classification of either positive (CM-positive sow) or negative (CM-negative sow) examples.

The classification was performed by starting from the root node until arriving at a leaf node. The descending order of the attributes within the decision-tree and the threshold values of the branches were calculated by the algorithm with the gain ratio criterion where the root of the tree represented the attribute with the highest information gain. In order to reduce the chance of overfitting, the C4.5-algorithm simplifies very highly and complex generated trees by the error-based pruning method.[14] The C4.5-algorithm is described in detail by Quinlan[14] and Mitchell.[15] The classification accuracy of the algorithm was tested with the stratified 10-fold cross-validation method which analyses the number of correctly and incorrectly classified instances (observations).[16] The whole dataset was randomly divided into ten subsets, nine parts being dedicated to train the algorithm and one for testing it. The training set was used by the algorithm for learning and building a decision tree and the test set was used to estimate the classification evaluation parameters. Then the algorithm ran ten times, each time with a different training and test set, and the results were validated. The classification performance assessment was evaluated with a two-dimensional confusion matrix consisting of the numbers of true positive (TP), false negative (FN), true negative (TN) and false positive (FP) classified examples. Sows with CM described the positive instances and healthy sows represented the negative instances in this study. The classification accuracy of the C4.5-algorithm was expressed by specific evaluation parameters (Table 2). The overall classification accuracy described the number of correctly classified instances in total. The proportion of correctly classified CM-positive sows in relation to all CM-positive sows was represented by the sensitivity. In addition, the specificity was defined by correctly classified CM-negative sows in relation to all CM-negative sows. The Kappa value reflected the degree of agreement for classifying the sows in the CM-positive or CM-negative classes. The error rate indicated the falsely classified CM-positive sows in proportion to all positively classified sows.

In this study, the minimum number of instances per outcome class varied between 20, 50 and 100, i.e. a new branch was created by the C4.5-algoithm only when it contained a number of instances greater or equal to the adjusted values of 20, 50 and 100. The results, calculated with the different minimum number of instances per class, were named according to the datasets A_{20}, A_{50}, A_{100} and B_{20}, B_{50}, B_{100}, respectively.

Table 2. Evaluation parameters of the classification accuracy of the C4.5-algorithm.

Evaluation parameters	Formula
Classification accuracy	TP+TN/(TN+FP+FN+TP) x 100
Sensitivity	TP/(TP+FN) x 100
Specificity	TN/(TN+FP) x 100
Kappa value	(TP+TN) - [((TP+FN) x (TP+FP) + (FP+TN) x (FN+TN))/N]/ N- [((TP+FN) x (TP+FP) + (FP+TN) x (FN+TN))/N] x 100
Error Rate	FP/(FP+TP) x 100

TP, true positive; TN, true negative; FP, false positive; FN, false negative, N, total number of instances.

Table 3. Evaluation parameters for Farms A (n = 987) and B (n = 596) with varied adjusted minimum number of instances per class

Dataset[a]	Classification accuracy	Sensitivity (%)	Specificity (%)	Error rate (%)	Kappa statistic (%)	No. of leaves	No. of nodes
A_{20}	53.2	54,7	48,4	46.4	6,4	5	9
A_{50}	54.2	55.9	47.5	45.4	8.4	4	7
A_{100}	55.0	58.1	48.2	44.8	10.0	4	7
B_{20}	61.2	65.8	44.2	36.3	21.7	8	15
B_{50}	60.2	65.5	46.0	37.4	19.6	4	7
B_{100}	56.4	64.0	52.6	41.1	11.5	3	5

a20, 50, 100 = at least 20, 50 or 100 instances per class.

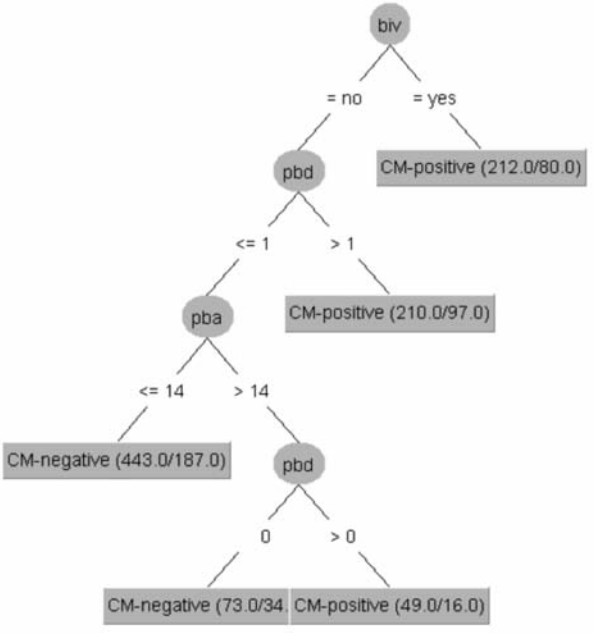

Figure 1. Decision tree showing the detected parameters and threshold values associated with CM of dataset A_{20} (n=987; minimum number of 20 instances per class); *biv*, birth intervention; *pbd*, piglets born dead; *pda*, piglets born alive.

Non-commercial use only

Results

The evaluation parameters varied between the two datasets and due to the specified number of instances per class (Table 3). The best values were achieved for Dataset A when the number of instances was set to the minimum of 100 instances per class and for Dataset B when the number of instances was set to the minimum of 20 instances per class. The evaluation parameters for B_{20} showed a better fit compared to A_{100}: The classification accuracy (61.2%) and the sensitivity (65.8%) of B_{20} were higher than for A_{100} (55.0%; 58.1%) and the error rate of B_{20} was 8.5% points lower. The Kappa value (21.7%) of B_{20} reached higher values compared to A_{100} (10.0%). The specificity of B_{20} (44.2%) was lower than for A_{100} (48.2%).

Graphical trees are presented for A20, A_{100}, B_{20} and B_{100} (Figures 1, 2, 3 and 4).

The decision-trees of both datasets showed differences, although the available attributes (*parity number, piglets born alive, piglets born dead, birth intervention, birth induction*) were the same for all trees

The attribute *birth induction* did not appear in any of the trees showing that the other parameters are more important for the occurrence of coliform mastitis. The attribute *parity number* was not chosen in the trees of Dataset A. The trees of A_{20}, A_{100} and B_{20} started with the attribute *birth intervention* as the root node which, therefore, was identified as the most influencing attribute.

In Dataset A_{20}, sows with no *birth intervention* but *piglets born dead* greater than zero and *piglets born alive* greater than 14 were CM-positive. In Dataset B_{20}, sows with no *birth intervention*, but a *parity number* less than or equal to three, *piglets born alive* greater than twelve and *piglets born dead* with at least one were CM-positive. The right sub-tree demonstrated that sows with birth intervention and *piglets born alive* greater than nine were CM-positive.

The decision-trees of A_{100} and B_{100} were pruned, which made the decision steps clearer and more generic. Therefore, attributes with a smaller information gain ratio were dropped by the algorithm; important parameters endured.

The tree size of A_{100} was decreased by one leaf and two nodes in comparison to A_{20}. The tree of B_{100} had five leaves and ten nodes less than B_{20}.

In Dataset A_{100}, sows were CM-positive when *birth intervention* was applied, with more than one *piglet born dead* or more than 14 piglets born alive. In Dataset B_{100}, sows with *piglets born alive* greater than ten and a *parity number* less than or equal to three were CM-positive.

Figure 2. Decision tree showing the detected parameters and threshold values associated with CM of dataset A_{100} (n=987; minimum number of 100 instances per class); *biv*, birth intervention; *pbd*, piglets born dead; *pda*, piglets born alive.

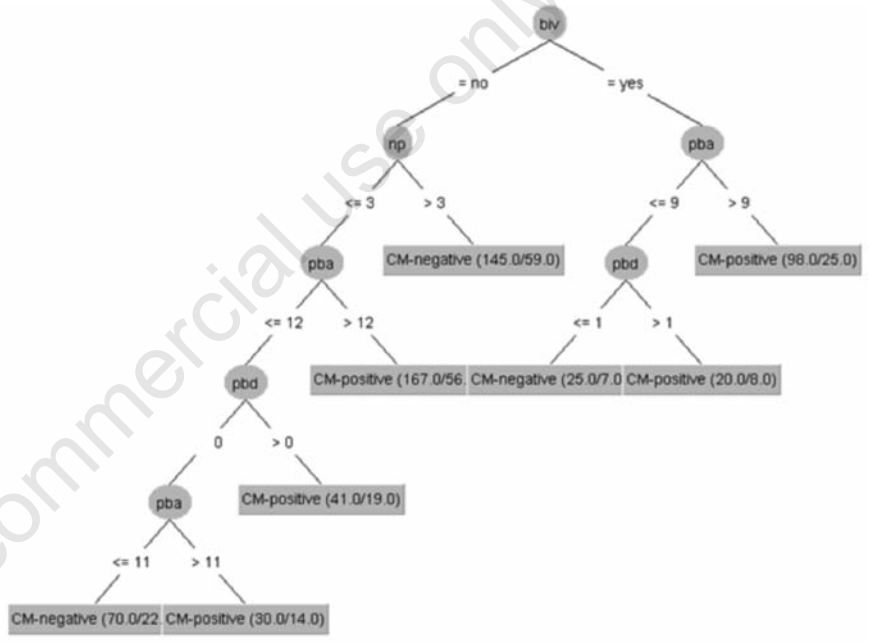

Figure 3. Decision tree showing the detected parameters and threshold values associated with CM of dataset B_{20} (n=596; minimum number of 20 instances per class); *biv*, birth intervention; *np*, parity number; *pbd*, piglets born dead; *pda*, piglets born alive.

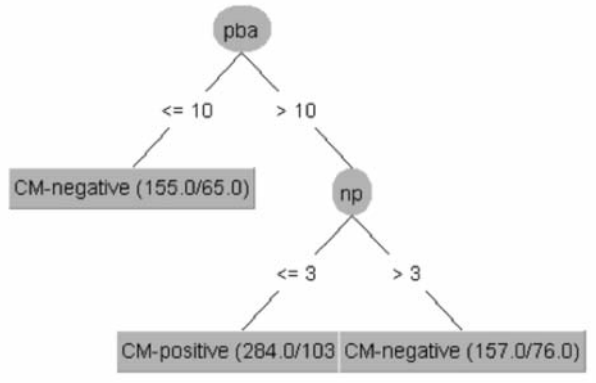

Figure 4. Decision tree showing the detected parameters and threshold values associated with CM of dataset B_{100} (n=596; minimum number of 100 instances per class); *biv*, birth intervention; *np*, parity number; *pbd*, piglets born dead; *pda*, piglets born alive.

Non-commercial use only

Discussion

The main objective of the study was the analysis of potential risk factors associated with sows suffering from coliform mastitis, determined on farm basis, by applying the C4.5-algorithm of the decision-tree technique. Environmental influences were standardised through the recording of data on these two farms with their high health standards.

According to the microbiological and genetic study design, only clear cases of CM-positive and selected cases of CM-negative sows were used for the analysis. Therefore, it is not possible to make statements of the real prevalence of CM on the farms where a large grey area of diseased sows exist.

The values of the evaluation parameters of the C4.5-algorithm were not acceptable compared to other studies. The sensitivity and specificity were too low and the error rate was too high. Kirchner *et al.* analysed culling strategies in swine breeding data by using the decision-tree technique and reached a classification accuracy value of about 85%.[10] The specificity was around 97% and the error rate on average 1%. Those datasets, however, consisted of 14,897 and 21,818 observations, much more than used in this study. Using more observations for model building improves the evaluation accuracy. With lower prevalence and therewith more skewed data, it is easier to reach higher accuracies.

The potential risk factors identified for CM by the decision-tree induction have also been described in other studies.[4,17,18] A higher number of piglets born alive was associated with a higher risk for the sows of becoming diseased. This is in accordance with findings of Bostedt *et al*.,[17] in which gilts with 1.1 piglets more than healthy sows suffered significantly more often from feverish puerperal illness, and also showed an increased stillbirth rate. Concerning the number of stillborn piglets, our study also supports these results. However, other researchers did not find differences in the number of stillborn piglets between diseased and healthy sows.[19,20]

Literature about the effect of the parity number on the occurrence of mastitis is contradictory. While Baer and Bilkei[21] found sows of higher parity (>4) having an increased risk of suffering mastitis, other studies have described a greater mastitis risk for lower parity sows (1. and 2. parity).[4,17,18] We also found a higher risk for primiparous sows, leading to the interpretation that those sows were more prone to disease. Explanations for this might be their not fully developed immune system,[22,23] or that sows suffering mastitis in their first parity might be culled. Physiological hyperthermia is also often observed in postparturient sows, especially primiparous ones, leading to misinterpretations.[7,24] The investigated factor birth intervention may be helpful in order to prevent CM in sows because it is associated with a higher risk for mastitis and can be regarded in the management. This fact has also been reported by Bostedt *et al*.[17] Manual intervention leading to a manipulation of the birth process might have a negative influence, especially if accompanied by insufficient hygiene.

In our study, the decision-tree technique was shown to have the ability to illustrate confirmed influencing factors for CM. In addition, the technique was able to weight those factors on farm basis. Individual herd and management differences were made clear by a different order of the attributes and different threshold values of the branches in the trees. Decision-trees, therefore, may allow exposure of individual weak points in the management of and comparisons between farms. In the context of multifactorial diseases, the utilisation of such a technique is shown feasible when certain conditions are fulfilled. For practical use, graphical trees should be smaller with clearly arranged decision steps to simplify interpretations for farmers and consultants. The minimum number of instances per branch has to be adjusted to the total number of instances, i.e. a small number of instances in total requires a small minimum number of instances per branch. The quality of the classification might be improved by optimising the study design and including more information about management and hygiene in the decision-tree algorithm.

Non-commercial use only

References

1. Bertschinger HU, Fairbrother JM. Escherichia coli infections. In: Straw BE, D'Allaire S, Mengeling WL, Taylor DJ (eds.), Diseases of swine. Iowa State University Press, Ames, IA, USA, 2006, pp. 431-68.
2. Bäckström L, Morkoc AC, Connor J, Larson R, Price W. Clinical study of mastitis-metritis-agalactia in sows in Illinois. J Am Vet Med Assoc 1984;185:70-3.
3. Hirsch AC, Philipp H, Kleemann R. Investigation on the efficacy of meloxicam in sows with mastitis-metritis-agalactia syndrome. J Vet Pharmacol Ther 2003;26: 355-60.
4. Krieter J, Presuhn U. Genetic variation for MMA treatment. Zuchtungskunde 2009;81: 149-54.
5. Awad Masalmeh M, Baumgartner W, Passering A, et al. Bakteriologische Untersuchungen bei an puerperaler Mastitis (MMA-Syndrom) erkrankten Sauen verschiedener Tierbestände Österreichs (Bacteriological studies in sows with pueperal mastitis in different herds in Austria). Tierarzt Umsch 1990;45:526-35.
6. Ross RF, Orning AP, Woods RD, et al. Bacteriologic study of sow agalactia. Am J Vet Res 1981;42:949-55.
7. Klopfenstein C, Farmer C, Martineau GP. Diseases of the mammary glands and lactation problems. In: Straw BE, Zimmermann JJ, Taylor DJ (eds.), Diseases of swine. Iowa State University Press, Ames, IA, USA, 2006, pp. 833-60.
8. Papadopoulos GA, Vanderhaeghe C, Janssens GPJ, et al. Risk factors associated with postpartum dysgalactia syndrome in sows. Vet J 2010;184:167-71.
9. van Asseldonk MAPM, Huirne RBM, Dijkhuizen AA, et al. Information needs and information technology on dairy farms. Comput Electron Agric 1999;22:97-107.
10. Kirchner K, Tolle KH, Krieter J. Decision-tree technique applied to pig farming datasets. Livest Sci 2004;90:191-200.
11. Gerjets I, Traulsen I, Reiners K, Kemper N. Comparison of virulence gene profiles of Escherichia coli isolates from sows with coliform mastitis and healthy sows. Vet Med 2010; (in press).
12. Furniss SJ. Measurement of rectal temperature to predict mastitis, metritis and alagactia (MMA) in sows after farrowing. Prev Vet Med 1987;5:133-9.
13. Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. ACM SIGKDD Expl Newsletter 2009;11: doi 10.1145/1656274.1656278.
14. Quinlan JR. C4.5: Programs for machine learning. 1993, Morgan Kaufmann, San Mateo, CA, USA.
15. Mitchell TM. Machine learning. 1997, McGraw Hill, New York, NY, USA.
16. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings 14th Int. Joint Conference on Artificial Intelligence, Montreal, Canada, 1995. Morgan Kaufmann, San Francisco, CA, USA.
17. Bostedt H, Maier G, Herfen K, Hospen R. Clinical examinations on gilts with pueperal septicaemia and toxaemia. Tierarztl Prax 1989;26;332-8.
18. Hoy S. Investigations on influence of different housing factors on frequency of puerperal diseases in sows. Prakt Tierarzt 2002;83:990-6.
19. Mirko CP, Bilkei G. Acute phase proteins, serum cortisol and preweaning litter performance in sows suffering from periparturient disease. Acta Vet Scand 2004;54:153-61.
20. Van Gelder KN, Bilkei G. The course of acute-phase proteins and serum cortisol in

mastitis metritis agalactia (MMA) of the sow and sow performance. Tijdschr Diergeneeskd 2005;130:38-41.
21. Baer C, Bilkei G. Ultrasonographic and gross pathological findings in the mammary glands of weaned sows having suffered recidiving mastitis metritis agalactia. Reprod Domest Anim 2005;40;544-7.
22. Wendt M. So optimieren Sie das Geburtsmanagement. Top Agrar 2000;1:6-8. (In German).
23. Hoy S. The impact of puerperal diseases in sows on their fertility and health up to next farrowing. Anim Sci 2006;82:701-4.
24. Gerjets I, Kruse S, Krieter J, Kemper N. Diagnosis of MMA affected sows: bacteriological differentiation, temperature measurement and water intake. Proceedings Int. Vet. Pig Soc. Congr., Durban, South Africa, 2008.

Non-commercial use only