Disease risk prediction based on an integrative data-methodological approach in dairy cattle

J. Lasser[1], C. Matzhold[2], C. Egger-Danner[1], F. Steininger[1], B. Fuerst-Waltl[4], T. Wittek[5] and P. Klimek[2]

[1]Graz University of Technology / Institute of Technical Informatics, Inffeldgasse 16/I, 8010 Graz / Complexity Science Hub Vienna, Josefstaedter Strasse 39, 1080 Vienna, Austria
[2]Medical University Vienna / Complexity Science Hub Vienna, Josefstaedter Strasse 39, 1080 Vienna, Austria
[3]ZuchtData EDV-Dienstleistungen GmbH, Dresdner Str. 89, 1200 Vienna, Austria
[4]Department of Sustainable Agricultural Systems, Division of Livestock Sciences, University of Natural Resources and Life Sciences (BOKU), Gregor-Mendel-Str. 33, 1180 Vienna, Austria
[5]Clinic of Ruminants, University of Veterinary Medicine, Veterinaerplatz 1, 1210 Vienna, Austria
Corresponding author: egger-danner@zuchtdata.at

Abstract

Cattle breeding has been data driven since decades ago, with central data storage and data processing. Breeding goals are including more and more trait complexes and more and more diverse data sources have been recorded over time. Pedigree and genomic information in combination with a variety of phenotypes for dairy, beef, functional and conformation traits are available. The increasing awareness and need for transparency in regard to health and welfare has led to further documentation and data sources as well as the further exploration of existing data sources, e.g. DHI with mid-infra-red spectra. The need to increase efficiency has raised interest in feeding and management information and technological advances are revolutionizing the dairy sector with a large amount of novel data sources. One of the aims in the D4Dairy project is to explore the joint use of diverse data in combination with machine learning methods disease risk prediction and prevention.

Based on information from 165 farms with data from different life domains such as milk recording, genetic, housing, nutrition, management, climate and health , algorithms to predict the disease risk for lameness, acute and chronic mastitis, anestrus, ovarian cysts, metritis, ketosis (hyperketonemia), and periparturient hypocalcemia (milk fever) have been derived. The results are encouraging, as, for example lameness can be predicted with high sensitivity and specificity (F1=0.74). Using three machine learning approaches of varying complexity (from logistic regression to gradient boosted trees) it was shown that to some extent the complexity of the algorithm can compensate for less diverse data. Presently the studies are ongoing with focus on elaboration of a data-driven decision support tool for early warning to reduce the disease risk.

Keywords: Big data, disease risk prediction, data integration.

Introduction

Due to technological progress a huge amount of different data is generated in dairy operations. Data and central data processing have been a main pillar for breeding for decades. Breeding based on data and central data processing has started more than fifty years ago. Data are generated from many different data sources on and off the farm. As breeding goals have expanded in recent decades, interest in linking various existing data sources has increased. These include traditional data from performance recording or linear scoring of animals, as well as veterinary diagnoses, data from claw trimming or data from laboratories. Since about 15 years ago, genomic information based on single-nucleotide-polymorphisms (SNP) is used in breeding and more and more genomic information is available. The advances from research on mid-infrared-spectra (MIR) show potential to use this information for early warning of diseases or optimization of feeding. The advances in sensor technology is another big step forward. More insight into animal physiology and behavior is possible, as well as more detailed information from the milking process. Feeding information, sensors measuring the housing climate in the stable as well as other information describing environmental conditions offer new possibilities for developing decision support tools to predict an animal's risk of disease.

Digitization offers many opportunities to develop new tools as well as to improve the existing approaches in high-efficiency dairy farming. Nevertheless, the challenge very often is the lack of standardization and linkage of data.

In the project D4Dairy, a highly integrated dataset with data from the various data sources mentioned above was created and a holistic approach for prediction of diseases was developed.

The research questions are based on the assumption that digitization has the potential to significantly improve early detection and prevention of animal diseases. With our research, we wanted to answer to main research questions

- Do complex integrated datasets enable new approaches to the detection of animal diseases or improve existing ones.
- Do advanced methodological approaches improve prediction performance?

Material and methods

The prerequisite to study these research questions is an integrated dataset fulfilling different requirements with regards to data quality and quantity as well as the methodological approaches for the analyses of large data sets.

The dataset used for this study was an existing data set from the project "Efficient Cow" (Egger-Danner *et al*., 2017) consisting of data from 166 farms, with 142 different variables, 6,519 cows and 45,944 observations. In addition, data from the national disease registry and national weather service were combined with the existing data set. The data set included data on diagnoses, housing, breed, age, management, conformation, feed, breeding values, lactation stage, environment and milk performance. The study concentrated on the eight most frequent diseases lameness, acute and chronic mastitis, anestrus, ovarian cysts, periparturient hypocalcemia, ketosis, metritis (Lasser *et al*. 2021).

In order to get more insight into the development of disease, as it is often an interaction of multiple factors, farm-risk-profiles were derived (Matzhold *et al*. 2021). A UMAP (Unifold Manifold Approximation and Projection) algorithm for dimension reduction and to identify pertinent factors was used. The HDBSCAN algorithm was used to identify clusters of farms with similar attributes. With multivariate regression models

the impact of individual risk factors was analysed. Each farm cluster identified a farm as a combination of environmental factors and management practices.

To predict the disease risk, three methods with increasing complexity were studied. The baseline was a logistic regression approach. In addition, two more complex machine learning approaches (random forests and gradient boosted trees – XGBoost) were employed to assess the improvements in disease prediction performance with increasing algorithm complexity. The methods were applied on two different datasets: A full dataset including information from the "Efficient Cow" project and weather information, and a reduced dataset that was comprised of data that is routinely collected by performance recording organisations.

Results and discussion

Figure 1 shows the farm-risk-profiles for the eight different diseases (Matzhold *et al*. 2021). Cluster 1 is characterized by the lowest prevalence of anestrus, ketosis, chronic mastitis, metritis and ovarian cysts. These are mainly farms in higher altitudes and with access to pasture. Cluster 4 showed the highest prevalence in anestrus, acute mastitis, lameness and ovarian cysts. These farms are characterized e.g. by less individualized feeding.

In Table 1 the F1-Score, Precision and Recall for predicting the eight diseases are shown for the logistic regression applied on the full dataset (Lasser *et al*., 2021). Precision is defined as probability that a predicted disease will actually be diagnosed, recall is the probability that an actual disease was correctly predicted. The F1-Score is the harmonic mean of precision and recall. Prediction works well for anestrus and lameness, with an F1-Score of 0.74. Prediction performance is worst for acute and chronic mastitis, with F1-Scores of 0.48 and 0.51, respectively. It has to be considered that the dataset collected did not include specific environmental and management factors related to mastitis e.g. hygiene measures, but it included specific information on housing and feeding with potential impact on lameness.

Table 2 shows the feature category importance for different diseases. The impact of the breeding values is rather small whereas environmental factors like feeding or housing show a high impact.

In Figure 2 precision and recall for the three different methods (logistic regression, random forest and XGBoost) and two datasets are presented. More complex methods perform better, especially when using the restricted dataset.

Conclusion and next steps

The studies on disease prediction using the "Efficient Cow" dataset showed promising potential of disease risk prediction using integrated datasets. The challenge remains the collection and integration of a data set with such diverse features.

In the next step the different approaches and methods will be applied on an even more complex and integrated dataset including additional data from daily milking and animal based sensors. The aim of the continued research is the development of a data-driven decision support tool providingearly warnings and enabling interventions before diseases fully emerge.

Figure 1. Farm-risk-profiles (Matzhold et al. 2021).

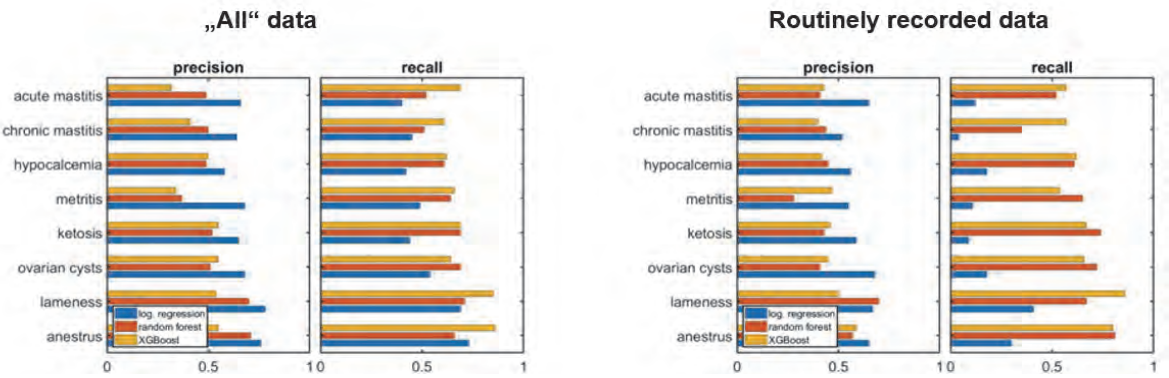

Figure 2: Precision and recall from different methods applied on 2 different datasets (Lasser et al. 2021).

Table 1. Logistic regression results (F1-Score, Precision, Recall) based on the "Efficient Cow" dataset (Lasser et al., 2021).

Disease	F1 Score	Precision	Recall
Anestrus	0,739	0,763	0,729
Lameness	0,737	0,780	0,700
Ovarian cysts	0,616	0,675	0,543
Ketosis	0,521	0,651	0,437
Metritis	0,549	0,677	0,490
Hypercalcemia	0,482	0,576	0,420
Chronic mastitis	0,514	0,635	0,445
Acute mastitis	0,479	0,656	0,395

Table 2. Feature category importance (Lasser et al., 2021).

Feature category	Anestrus	Lameness	Ovarian cysts	Ketosis	Metritis	Periparturient hypocalcemia	Chronic mastitis	Acute mastitis
Age	4,9	36,7	2,8	4,7	2,2	28,4	7,1	11,3
Breed	1,8	2,6	1,2	1,5	2,4	0,4	0,3	2,5
Breeding values	3.0	2,4	2,7	1,7	2,7	1,9	4.0	2,5
Diagnosis source	16.0	15,6	29,6	11,9	33,9	7,1	18,9	16,3
Environment	14,3	4,9	16,5	10,5	15,6	13,8	19,8	10,1
Feed	27,4	11,8	17,5	22,4	12,7	7,5	18.0	19,9
Housing	11,5	12,9	11,3	19.0	15,6	5.0	7,4	13,6
Husbandry	9,1	5.0	4,5	8,8	7.0	2,7	6,4	6,4
Lactation stage	2,7	1,2	4,7	16,1	1,8	28,9	0,7	1,3
Milk indicators	5,9	2.0	3,6	0,7	2,2	1,4	10,4	12,5
Physical indicators	3,6	4,8	5,7	2,7	3,9	3.0	7,1	3,7

[1]Cumulative permutation feature importance contributions for the eleven feature categories. Values are given in % of the sum of all feature importances for a given disease.

Acknowledgement

This work was conducted within the COMET-Project D4Dairy (Digitalisation, Data integration, Detection and Decision support in Dairying, Project number: 872039) that is supported by the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), the Austrian Federal Ministry for Digital and Economic Affairs (BMDW) and the provinces of Lower Austria and Vienna in the framework of COMET-Competence Centers for Excellent Technologies. The COMET program is handled by the FFG.

References

Egger-Danner, C., B. Fuerst-Waltl, C. Fuerst, L. Gruber, S. Hörtenhuber, A. Köck, M. Ledinek, C. Pfeiffer, F. Steininger, R. Weissensteiner, *et al***.** (2017) Efficient Cow-Analyse und Optimierung der Produktionseffizienz und der Umweltwirkung in der Österreichischen Rinderwirtschaft In: Abschlussbericht. Tech. Report. Wien: Zentrale Arbeitsgemeinschaft Österreichischer Rinderzüchter (ZAR). Available from https://raumberg-gumpenstein.at/ forschung/infothek/downloads/ download.html?path=Forschungsberichte%252F12_2017_ gruber_efficient_cow_ abschlussbericht.pdf [accessed November 3, 2021].

Lasser, J., Matzhold, C., Egger-Danner, C., Fuerst-Waltl, B., Steininger, F., Wittek, T., and Klimek, P. (2021) Integrating diverse data sources to predict disease risk in dairy cattle – a machine learning approach. Journal of Animal Science 99(11), skab294.

Matzhold, C., Lasser, J., Egger-Danner, C., Fuerst-Waltl, B., Wittek, T., Kofler, J., Steininger, F., and Klimek, P. (2021) A systematic approach to analyse the impact of farm-profiles on bovine health. Science Reports 11, 21152. doi: 10.1038/s41598-021-00469-2