Machine Learning/AI/Data Analysis - what is necessary to make a DVM spend time with data analysis today?

Author

Doerte Doepfer, DVM, MSc, PhD, professor, Department of Medical Sciences, School of Veterinary Medicine, UW-Madison, dopfer@wisc.edu

Introduction

Food animal practices are increasingly involved in receiving and processing data, for example from herd management programs, to derive evidence-based decisions for their consultancy work. Precision medicine and artificial intelligence (AI) -enhanced tools for the generation of data, data editing, and the analysis of data are here to stay. While DVM's are food animal disease, prevention, and production experts, they are often not trained to analyze such complex data streams,and one of the reasons may be because they 'did not sign up for this'. With the focus on AI-enhanced and multiple machine learning tools available to food animal practice, deriving decisions and analysis from data streams is more than 'hiring some computer scientist' to do it. A DVM consultant's time should be increasingly saleable for the value that veterinary expertise brings to deriving information from data sets andfor making predictions about life outcomes of food animals among which cattle.

'Good to know' about applied stats…

Data sets are valuable assets for informed decision-making processes, and an understanding of applied statistics shaped into a basic toolbox benefits the quantification of associations between risk factors and outcomes of disease, production, longevity, welfare and many others.The following is a beginner's walk-through a data set and its very basic statistical analysis. For example, take a data set of n=2741 fresh dates collected during 5 years on a ~1200 cow commercial dairy in the Midwest of the US, start with a cross table of frequencies in addition to a bar chart showing the proportion from the same cross table of an outcome(Removed<=60DIM) stratified by a risk factor (displaced abomasum: DA<=60DIM yes or no). Figure 1 shows such a cross table and bar chart of proportions generated using the Pivot table tool in Windows Excel(Free Online Spreadsheet Software: Excel | Microsoft 365).There are already two challenges listed above: obtaining the data set formatted for making a pivot table and generating the pivot table and bar chartusing Microsoft software.

(a)

	Outcome:	**Removed <=60DIM**			
Exposure:		y	n		
DA<=60DIM	y	14	38	52	N1
	n	115	2574	2689	N0
		129	2612	2741	
		M1	M0		

(b)

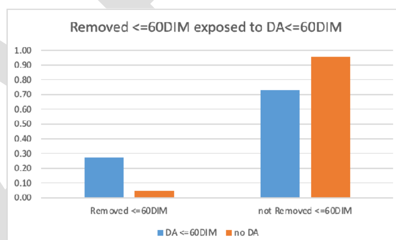

Figure 1: cross table of frequencies (a) and bar chart of proportions for the outcome Removed<=60DIM for having DA<=60DIM from n=2741 fresh dates collected during 5 years on a commercial midwestern dairy farm (~1200 cows)

Calculating the Relative Risk (RR) and its 95% Confidence interval (95% CI) of being removed <=60DIM while being exposed to a DA<=60DIM can be accomplished by dividing the incidence in the exposed (a/(a+b)) over the incidence of the non-exposed (c/(c+d)) as shown in Figure 2.The resulting RR indicates that a cow being removed <=60DIM while exposed to a DA<=60DIM is 6.30 (3.89 – 10.20 95% CI) times more frequent compared to being removed and not having had a DA. Similar calculations are completed for the Odds Ratio ((a*d)/(b*c)) and its 95% confidence interval. Confidence interval limits are approximated using time series as long as frequencies in the two-by-two table are >5. A fisher's exact p-value is used to evaluate whether the RR is statistically significantly smaller than 0.05 whenever one of the 4 cells in the two-by-two table is 5 or less (see https://www.graphpad.com/quickcalcs/contingency1/). Please note that the outcome is on top of the table as column label, and the potential risk factor is the row label. Whenever the RR or OR are >1, one concludes that a risk factor is at play while if <1, a protective factor has been identified. As soon as RR=1 (RR=1 means equal frequency between exposed and non-exposed) is included in the confidence interval, one cannot reject the H0 (null hypothesis) that RR is statistically significantly different from 1 at the 95% confidence level. A spreadsheet to do these calculations is available from the author upon request.

Iexp	0.27	a/(a+b)			
Inonexp	0.04	c/(c+d)			
RR	6.30	[a/(a+b)]/[c/(c+d)]	OR	8.25	(a*d)/(b*c)
LN(RR)	1.84		LN(OR)	2.11	
VAR(LN(RR))	0.06	[b/(a*(a+b))]+[d/(c*(c+d))]	VAR(LN(OR	0.11	(1/a)+(1/b)+(1/c)+(1/d)
SD(LN(RR))	0.25	VAR(LN(RR))^0.5	SD(LN(OR))	0.33	square root of the variance
Lower95%CI	3.89	EXP(LN(RR)-(1.96*SD(LN(RR)))	Lower95%(	4.35	EXP(LN(OR)-(1.96*SD(LN(OR)))
High95%CI	10.20	EXP(LN(RR)+(1.96*SD(LN(RR)))	High95%CI	15.65	EXP(LN(OR)+(1.96*SD(LN(OR)))
Z95%	1.96		Z95%	1.96	
Z90%	1.61		Z90%	1.61	

remember:			
y	n		
a	b	a+b	N1
c	d	c+d	N0
a+c	b+d		
M1	M0		

Figure 2: Relative risk (RR) and Odds ratio (OR) and 95% Confidence interval (95% CI) for being Removed <=60DIM while exposed to a displaced abomasum: DA<=60DIM; data from n=2741 fresh dates collected during 5 years on a commercial midwestern dairy farm (~1200 cows); Z95% and Z90% are the Z-values used to calculate the confidence interval limits, VAR: variance, EXP: exponent e^x

These calculations seem tedious in this context but this type of so-called 'measures of association' allow an insight into quantity and direction of the association between outcome and risk factor in addition to having a connection to many of the ML algorithms the modern AI-world is exposing us to. For example, a simple regression is shown as an

Equation #1: $y = a + b*x + error$,

where y is the outcome and b is the risk factor, also called predictor (a is the intercept, that is the point of the y-axis where the regression line crosses). When exponentiating the slope estimate of the regression line 'b', one obtains the OR from above. This is a simple corss-talk between two-by-two tables and logistic regressions (Jr and Lemeshow, 2004).

How can we generate the data needed to produce Figures 1 and 2? For example, the DairyCoPilot interactive application is a web-based tool (link: https://connect.doit.wisc.edu/dairy-copilot/) that receives an event list from DairyComp305 (Dairy Herd Management Program - DairyComp | VAS), edits and cleans it for direct export into Microsoft Excel or for automated data analysis ranging from descriptive to bivariate and multiple variable data analysis(Aravamuthan et al., 2024). The cited article by Aravamuthan et al. 2024 provides access to DairyCoPilot, the data set used during this exercise, and the data analysis tools with instructions. The tool is free for use.

In addition, the resulting regression Equation #1 can be used to make prediction models for the outcome (y) using additional computations. This basic quantification of an association using a simple regression is the base for many Machine learning (ML) algorithms among which Artificial Neural Networks (ANN). Manysuch regression models together can be composed into large neural networks that generate AI-enhanced predictions. Understanding these baseline associations is a first step to gaining insight into large-scale models of high complexity(Chollet, 2018).

The process of quantifying associations between risk factors and outcomes can be scaled up into large and Big data analysis engines among which, for example,precision dairy medicine. Example for such data analysis engines and efforts are: MmmooOgle(MmmooOgle for dairy farmers passionate about performance), the Holstein Association USA from where Dr J Bewley is overseeing the WKU SmartHolstein Lab(WKU SmartHolstein Lab), and as a final example, the Dairy Innovation Hub at UW-Madison in Wisconsin(Dairy Innovation Hub – Agricultural & Life Sciences).

ChatGPT, Gemini and alike for data analysis…

Instead of going through the steps of statistical data analysis, the use of Large Language Models (LLM), for example using ChatGPT (ChatGPT)or Gemini(Gemini - chat to supercharge your ideas) among others, for deriving information from data sets is increasing rapidly. One can simply upload an excel file and ask the engine to analyze the data set for risk factors associated with an outcome of choiceand obtain the source of information (ALWAYS CHECK OUTCOMES AND SOURCES FOR CORRECTNESS!!!!). In addition, one can ask for the software code to perform the data analysisand for a text summary of the outcomes meant for a specialized audience of stakeholders. It is questionable to submit data owned by producers to a public website where it becomes public domain though. Is it acceptable to use the top 3 rows of data together with the explanation of variables to prompt ChatGPT for a pseudo-data analysis? Such LLM interactive tools are made to please the user, and therefore the engine produces the response and graph of a network of risk factors related to an outcome like shown in Figure 3 and Appendix 1. Fine-tuning expert systems based on LLM applications can contribute to data analysis tasks and interactive chatbots that play the roles of dairy and beef consultants. Critical evaluation of the outcomes is key because the models are known to 'hallucinate', discriminate, and sometimes they produce errors. Since LLM's are a phenomenon of our time, it is important to know about these tools in the author's opinion. While generating such expert systems, it is better to have experts present to scrutinize the set-up and outcomes. Therefore, a basic understanding of such tools is a useful time investment.

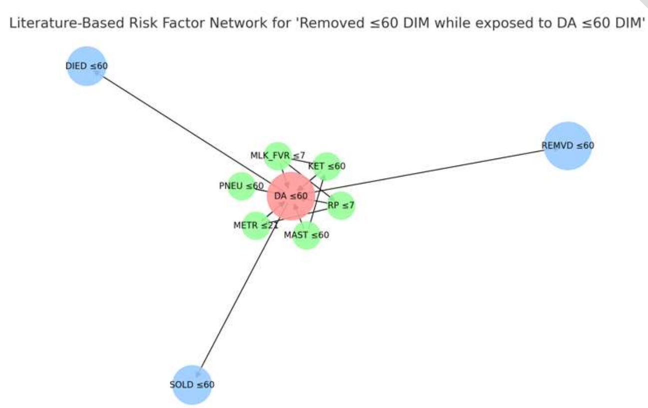

Figure 3: an output from ChatGPT prompted by the list of variables from the dairy production data set processed through DairyCoPilot. Part of the text output is shown in Appendix 1.

Conclusions

Cybersecurity, data safety, data privacy, and data ownership are discussion points for data sets processed like the one from the DairyCoPilot example. We will learn more about the consequences of Generative AI, Persuasive AI, and Degenerative AI (AI predictions generated from data sets simulated by AI models themselves) for dairy and beef cattle consulting and for informed decision-making processes in the near future.

References

Aravamuthan, S., D. Dopfer, and E. Walleser. 2024. DairyCoPilot-Automated data compilation and analysis tools for DairyComp data assets. PLoS One. 19:e0297827. doi:10.1371/journal.pone.0297827.

ChatGPT. Available from: https://chat.openai.com

Chollet, F. 2018. Deep Learning with Python. Manning Publications Co, Shelter ISland, NJ, USA.

Dairy Innovation Hub – Agricultural & Life Sciences. Available from: https://cals.wisc.edu/dairy-innovation-hub/

Free Online Spreadsheet Software: Excel | Microsoft 365. Available from: https://www.microsoft.com/en-us/microsoft-365/excel

Gemini - chat to supercharge your ideas. Gemini. Available from: https://gemini.google.com

Hosmer Jr, D. W. H., and S. Lemeshow. 2004. Applied Logistic Regression. John Wiley & Sons.

MmmooOgle for dairy farmers passionate about performance. MmmooOgle. Available from: https://www.mmmooogle.com/?lang=nl

WKU SmartHolstein Lab. WKU SmartHolstein Lab. Available from: https://www.smartholstein.com

Appendix 1: ChatGPT output prompted using the list of variables in the dairy production data set from DairyCoPilotand explanations about them:

Summary of Findings for Dairy Producers

(Note: this text is written by ChatGPT, the author of this abstract is not responsible for this text!)

As a dairy producer, understanding the factors that lead to early cow removal (within the first 60 days in milk, or DIM) is crucial for maintaining herd health and optimizing productivity. The following summary highlights key findings related to early removal due to displaced abomasum (DA) and associated risk factors:

1. **Displaced Abomasum (DA) as a Central Issue**: DA within the first 60 DIM is a significant condition that can lead to cows being removed from the herd either through sale or death. DA is often a result of metabolic imbalances, which are common during the early lactation period.
2. **Metabolic Disorders**: Conditions such as ketosis, milk fever, and retained placenta are closely linked to DA. Cows that experience these metabolic disorders are at a higher risk of developing DA and consequently being removed from the herd.
3. **Reproductive Health**: Uterine infections like metritis and complications related to retained placenta can predispose cows to DA. Ensuring good reproductive health management, especially during the transition period, is key to reducing the incidence of DA.

4. **Infectious Diseases**: Mastitis and pneumonia are significant health challenges in early lactation. These diseases can weaken cows and increase their susceptibility to DA, leading to early removal. Prompt detection and treatment of these conditions are essential.
5. **Early Lactation Monitoring**: The first 60 days post-calving is a critical period where cows are most vulnerable to these health issues. Regular monitoring of body condition, metabolic indicators, and overall health can help identify at-risk cows early and allow for timely interventions.
6. **Preventive Management Strategies**: Implementing effective transition cow management, including nutritional support to prevent metabolic diseases, proper calving practices, and close monitoring for early signs of illness, can significantly reduce the risk of DA and early removal.
7. **Economic Impact**: Early removal of cows due to health issues not only affects herd productivity but also incurs financial losses. Investing in preventive measures and early intervention strategies will save costs associated with cow replacement and lost milk production.

Actionable Recommendations

- **Nutritional Management**: Ensure that cows receive a balanced diet, particularly during the transition period, to prevent metabolic disorders like ketosis and milk fever.
- **Health Monitoring**: Implement routine health checks, including monitoring for signs of DA, mastitis, pneumonia, and other disorders that can lead to early removal.
- **Reproductive Health**: Focus on reproductive health by managing calving ease, preventing retained placenta, and addressing any post-calving complications promptly.

- **Staff Training**: Educate farm staff on recognizing early symptoms of health issues and the importance of early intervention to prevent escalation.

By focusing on these areas, dairy producers can reduce the incidence of early cow removal, improve overall herd health, and enhance farm profitability.