Interuniversity Master
in Statistics and Operations Research UPC-UB

Master's Thesis

A compositional study of biochemical and haematological factors involved in calf lameness

Mona Thiele

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

Feedlot Farming in Numbers

https://ourworldindata.org/meat-production

- Since 1961 the worldwide per capita meat consumption has increased around 20 kilograms per year.
- In the same time beef production has more than doubled in size.

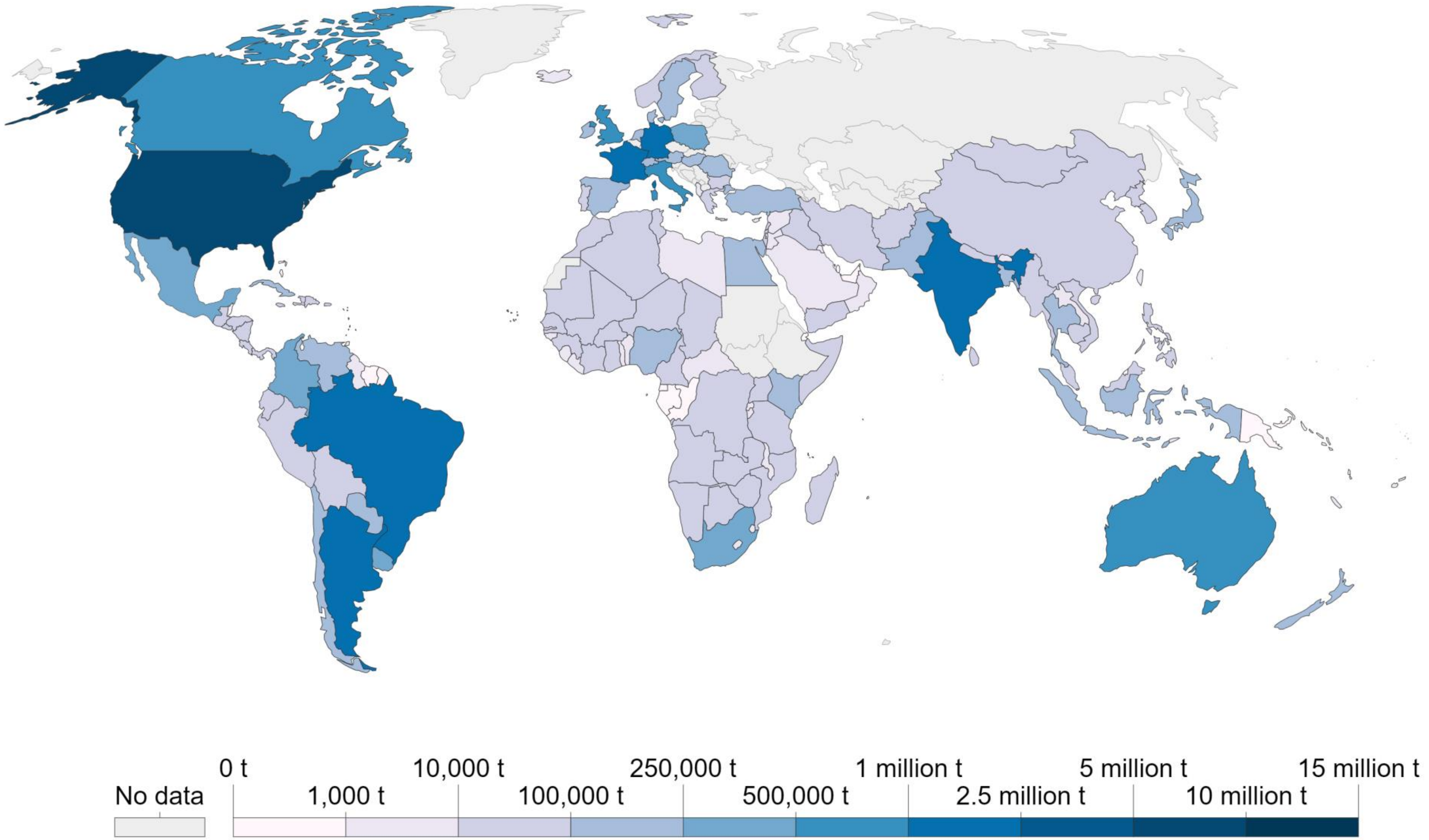
www.statista.com

- The average person in Spain consumed 4.9 kilos of beef per year in 2019.

www.farmonline.com.au

Beef production, 1961

Our World in Data

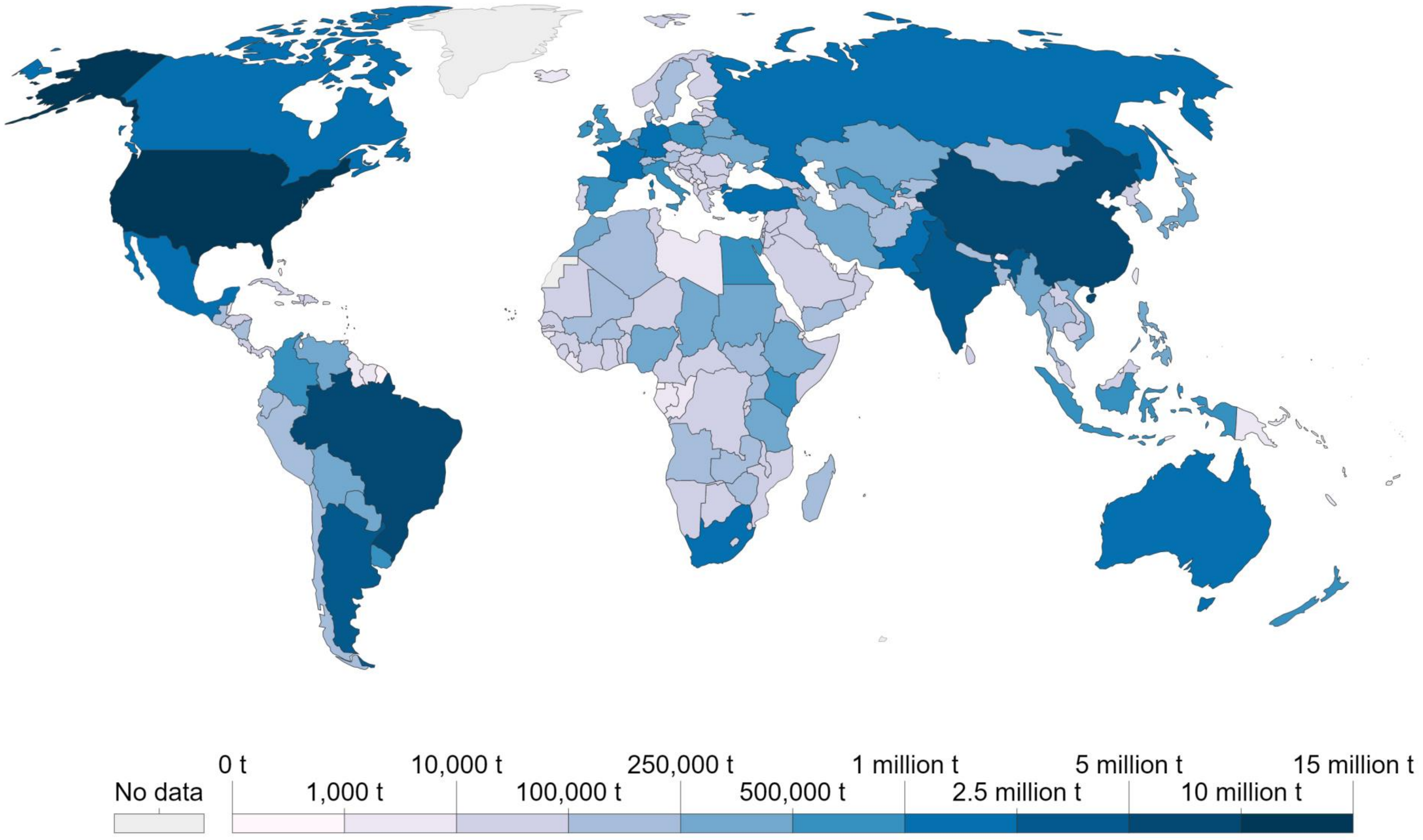


Source: UN Food and Agricultural Organization (FAO)

Note: Beef and buffalo (cattle) meat production from both commercial and farm slaughter. Data are given in terms of dressed carcass weight, excluding offal and slaughter fats.

Beef production, 2018

Our World in Data



Source: UN Food and Agricultural Organization (FAO)

Note: Beef and buffalo (cattle) meat production from both commercial and farm slaughter. Data are given in terms of dressed carcass weight, excluding offal and slaughter fats.

Feedlot Farming in Numbers

Consequences of this development:

- Cattle kept in dense herds of large numbers.
- Increased exposure to stress.
- Infectious diseases spread faster due to proximity of animals and often poor hygiene of feedlots.
- Preventive feeding of antibiotics on grand scale lead to antibiotic resistance in cattle and humans.

Lameness in cattle

- Lameness prevalence in cattle population has been reported to be as high as 36.8% (Shearer et al, 2013).
- Comprised of a group of infectious and non-infectious diseases, as well as different injuries.
- Impairment of the walking ability of one extremity or more.
- Various degrees of pain and stress.

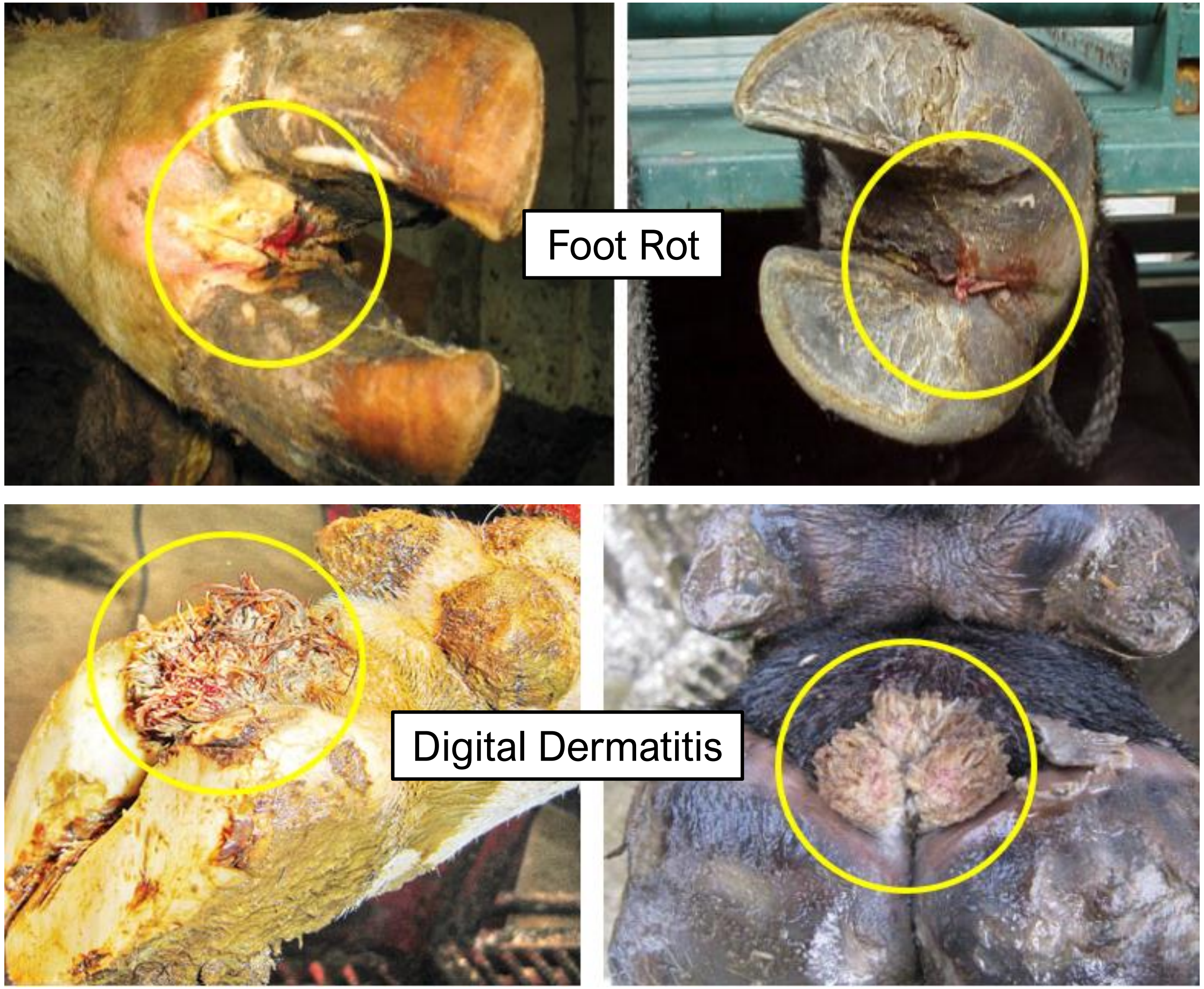
Score	Lameness	Signs
1	Sound	
2	Mild	Stands with flat back, but arches when walks. Gait is slightly abnormal.
3	Moderate	Stands and walks with arched back. Moves with short strides; reduced weight bearing can be detected on affected leg. Head drops when weight is taken on affected leg.
4	Severe	Back arched when standing and walking, obvious reduced weight bearing on affected limb. Cow moves slowly, often making frequent stops, and may show secondary signs of pain such as weight loss, teeth grinding and excess salivation.
5	Highly severe	Back arched, reluctant to move. Does not bear weight on the affected leg.

Lameness score: visual characterization of severity of lameness in feedlot cattle and the signs used for categorization (modified from Sprecher et al. (1997)).

Lameness in cattle

- Lameness prevalence in cattle population has been reported to be as high as 36.8% (Shearer et al, 2013).
- Comprised of a group of infectious and non-infectious diseases, as well as different injuries.
- Impairment of the walking ability of one extremity or more.
- Various degrees of pain and stress.

Adapted from Currin et al (2015)



https://www.progressivedairy.com/
https://sites.google.com/site/digitaldermatitisnz/

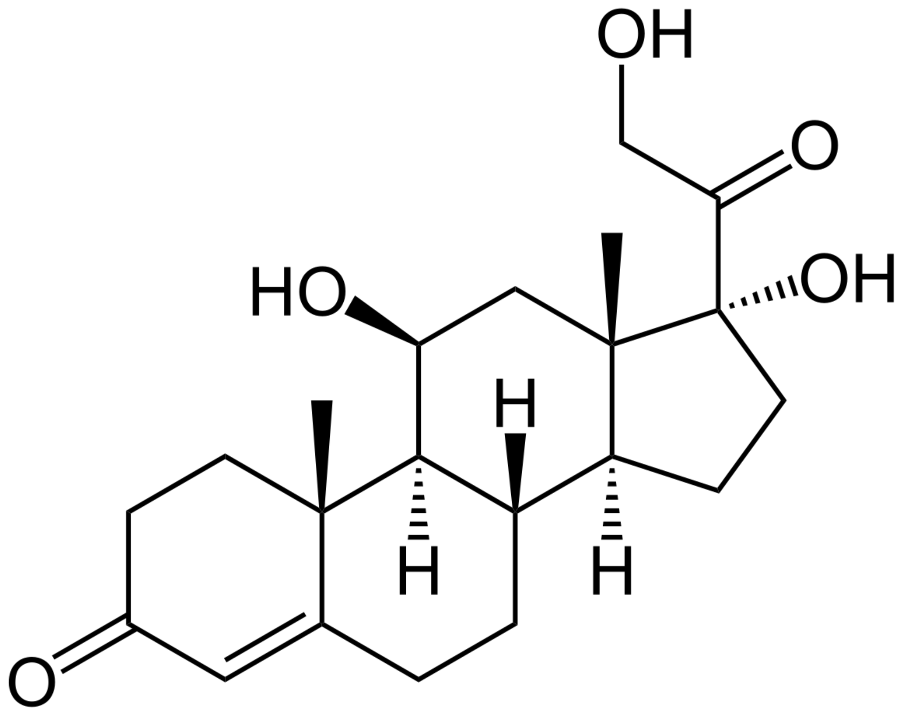
Biochemical indicators for stress and inflammation

Cortisol

- One the first indicators of pain that has been connected to the lameness score.
- Blood cortisol proved very unstable in various studies.
- Possible alternatives:

  Hair Cortisol

  Substance P



https://en.wikipedia.org/wiki/Cortisol

Biochemical indicators for stress and inflammation

Cortisol

- One the first indicators of pain that has been connected to the lameness score.
- Blood cortisol proved very unstable in various studies.
- Possible alternatives:

  Hair Cortisol

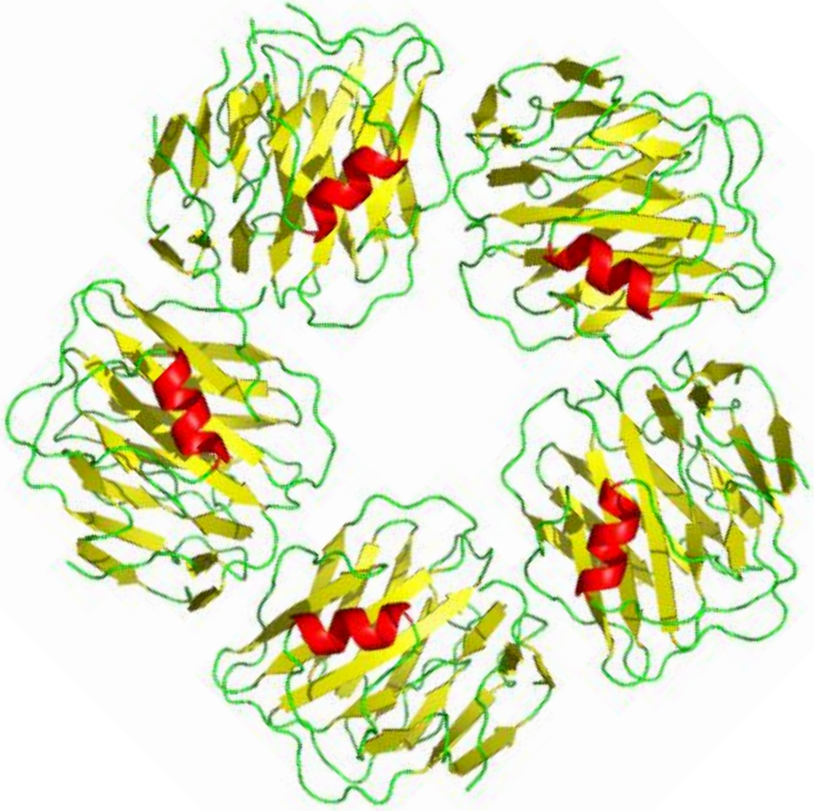
  Substance P

Acute phase proteins (APPs)

- Linked to the acute phase response after tissue injury
- triggering the immune reaction to the injury: Inflammation
- Two important APPs:

  Haptoglobin

  Serum amyloid A (SAA)



https://en.wikipedia.org/wiki/Serum_amyloid_P_component

Biochemical indicators for stress and inflammation

Cortisol

- One the first indicators of pain that has been connected to the lameness score.
- Blood cortisol proved very unstable in various studies.
- Possible alternatives:

  Hair Cortisol

  Substance P

Acute phase proteins (APPs)

- Linked to the acute phase response after tissue injury
- triggering the immune reaction to the injury: Inflammation
- Two important APPs:

  Haptoglobin

  serum amyloid A (SAA)

Further indicators

- Rectal temperature (fever)
- Different haematological values:

  Leucocyte number

  Percentage of neutrophils

  White blood cell count

Data set on calf lameness

- Data set on calf lameness provided by IRTA researcher Sònia Marti.
- Over 1300 observations from individual animals and their haematological and biochemical measures.

Categorical variables

Physiological, biochemical and haematological variables

Sample	Lesion	Lameness	Severity	Rectaltemp	SAA	SubP	Hapto	Hair	NL	Cortisol	RBC	MCV	HCT	PLT	MPV	HGB	WBC	LYM	MONO	GRAN
1	3	4	2	103.6	301.3	16.1	2.2	.	1.3	3.7	7.4	41.7	0.3	234.0	5.7	10.8	11.4	4.3	1.3	5.8
2	7	3	1	99.1	194.1	21.4	0.1	.	1.1	4.8	7.7	39.1	0.3	45.0	0.0	10.7	6.4	2.9	0.4	3.1
3	1	4	2	104.1	107.2	25.9	4.4	.	2.1	1.8	7.9	39.6	0.3	370.0	5.7	11.1	12.8	3.7	1.2	7.9
4	7	1	0	104.6	136.6	12.5	0.3	.	1.9	11.6	7.9	41.2	0.3	263.0	5.9	11.5	9.7	3.0	0.9	5.8
6	7	3	1	104.5	206.7	16.3	1.2	.	.	2.0	.	.	.	.	.	13.0	.	.	.	.
7	3	3	1	101.9	256.8	14.9	4.2	.	0.5	3.3	6.7	39.2	0.3	125.0	5.9	10.0	7.7	4.8	0.6	2.3
8	7	2	1	103.4	178.9	32.5	0.1	.	0.5	3.5	3.1	37.0	0.1	94.0	5.7	4.4	2.3	1.3	0.3	0.7
9	1	4	2	103.7	219.1	16.5	4.8	.	1.7	6.5	8.8	37.6	0.3	435.0	5.4	12.4	12.0	4.0	1.1	6.9
10	7	2	1	103.4	166.9	28.2	0.2	.	1.2	1.5	9.8	41.0	0.4	337.0	5.8	14.3	11.6	4.7	1.1	5.8
11	4	2	1	102.8	123.1	16.3	1.4	.	1.0	5.8	7.8	41.0	0.3	492.0	5.5	12.0	7.4	3.3	0.7	3.4

Data set on calf lameness

- Data set on calf lameness provided by IRTA researcher Sònia Marti.
- Over 1300 observations from individual animals and their haematological and biochemical measures.
- Lesion causing lameness diagnosed by veterinarian.

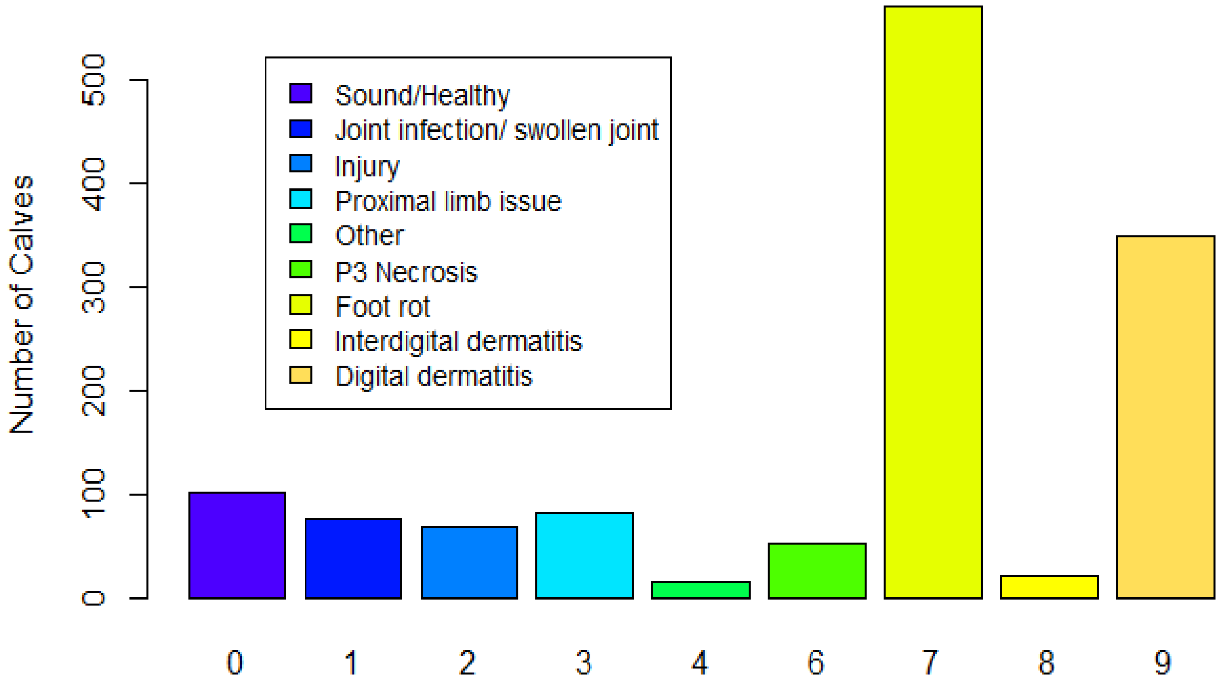
Categorical variables

Physiological, biochemical and haematological variables

Sample	Lesion	Lameness	Severity	Rectaltemp	SAA	SubP	Hapto	Hair	NL	Cortisol	RBC	MCV	HCT	PLT	MPV	HGB	WBC	LYM	MONO	GRAN
1	3	4	2	103.6	301.3	16.1	2.2	.	1.3	3.7	7.4	41.7	0.3	234.0	5.7	10.8	11.4	4.3	1.3	5.8
2	7	3	1	99.1	194.1	21.4	0.1	.	1.1	4.8	7.7	39.1	0.3	45.0	0.0	10.7	6.4	2.9	0.4	3.1
3	1	4	2	104.1	107.2	25.9	4.4	.	2.1	1.8	7.9	39.6	0.3	370.0	5.7	11.1	12.8	3.7	1.2	7.9
4	7	1	0	104.6	136.6	12.5	0.3	.	1.9	11.6	7.9	41.2	0.3	263.0	5.9	11.5	9.7	3.0	0.9	5.8
6	7	3	1	104.5	206.7	16.3	1.2	.	.	2.0	.	.	.	.	.	13.0	.	.	.	.
7	3	3	1	101.9	256.8	14.9	4.2	.	0.5	3.3	6.7	39.2	0.3	125.0	5.9	10.0	7.7	4.8	0.6	2.3
8	7	2	1	103.4	178.9	32.5	0.1	.	0.5	3.5	3.1	37.0	0.1	94.0	5.7	4.4	2.3	1.3	0.3	0.7
9	1	4	2	103.7	219.1	16.5	4.8	.	1.7	6.5	8.8	37.6	0.3	435.0	5.4	12.4	12.0	4.0	1.1	6.9
10	7	2	1	103.4	166.9	28.2	0.2	.	1.2	1.5	9.8	41.0	0.4	337.0	5.8	14.3	11.6	4.7	1.1	5.8
11	4	2	1	102.8	123.1	16.3	1.4	.	1.0	5.8	7.8	41.0	0.3	492.0	5.5	12.0	7.4	3.3	0.7	3.4

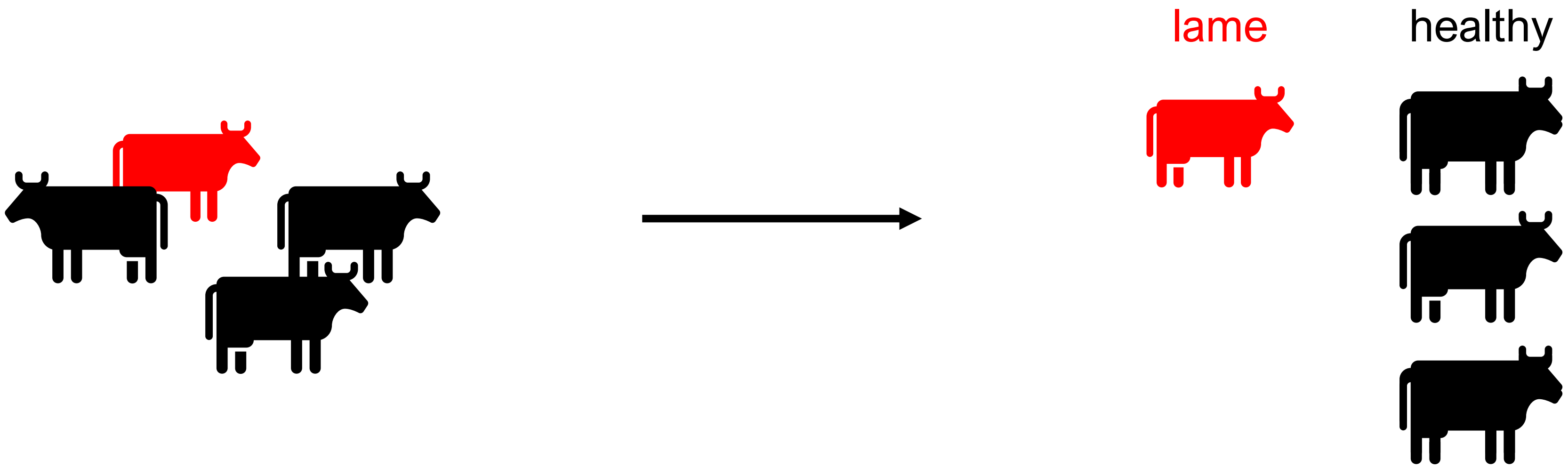
Data set on calf lameness

- Data set on calf lameness provided by IRTA researcher Sònia Marti.
- Over 1300 observations from individual animals and their haematological and biochemical measures.
- Lesion causing lameness diagnosed by veterinarian.



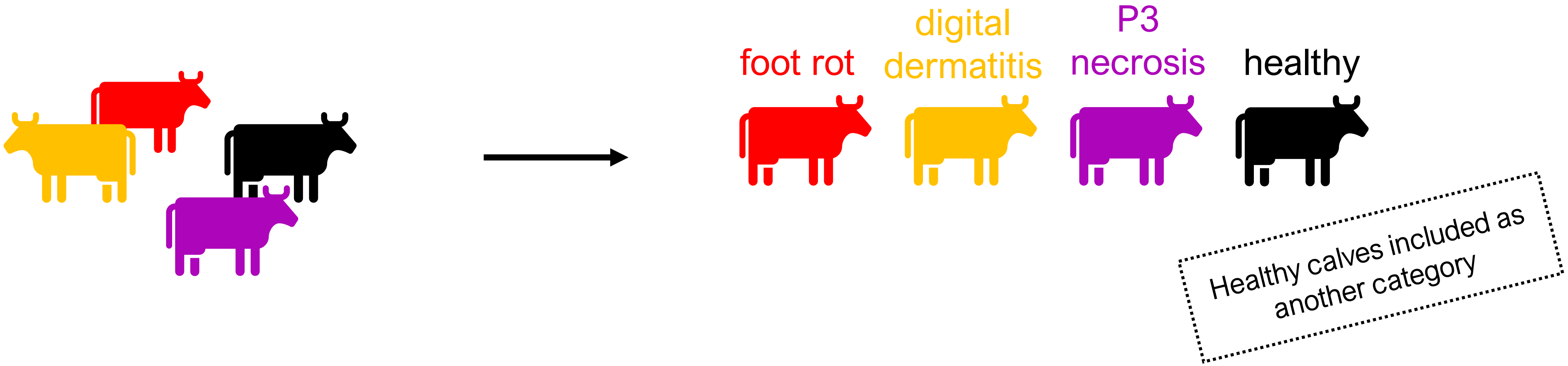
Research Questions

Can we distinguish lame and healthy calves by analysing their blood work?



Research Questions

Given lameness, can we predict the most probable lesion causing this condition using haematological and biochemical data?



Data processing and analysis

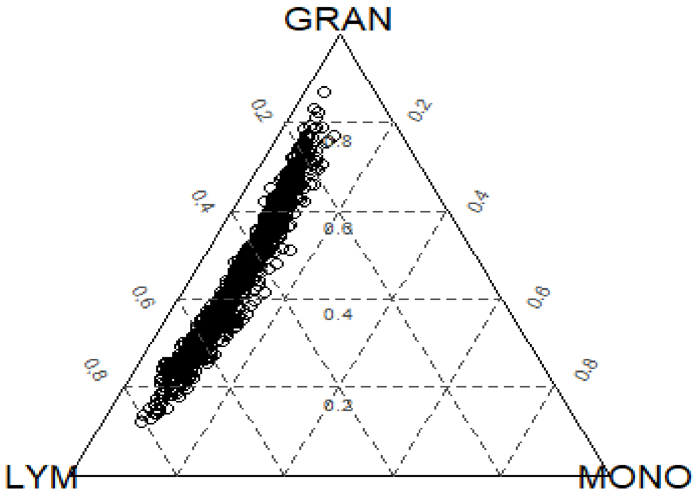
<u>Data processing:</u>

- Square root or logarithmic transformation of appropriate variables.
- Compositional haematological data: isometric log-ratio transformation.

$$WBC = GRAN + MONO + LYM$$

$$\boldsymbol{Ilr1} = \frac{1}{\sqrt{2}} * \log\left(\frac{GRAN}{LYM}\right)$$

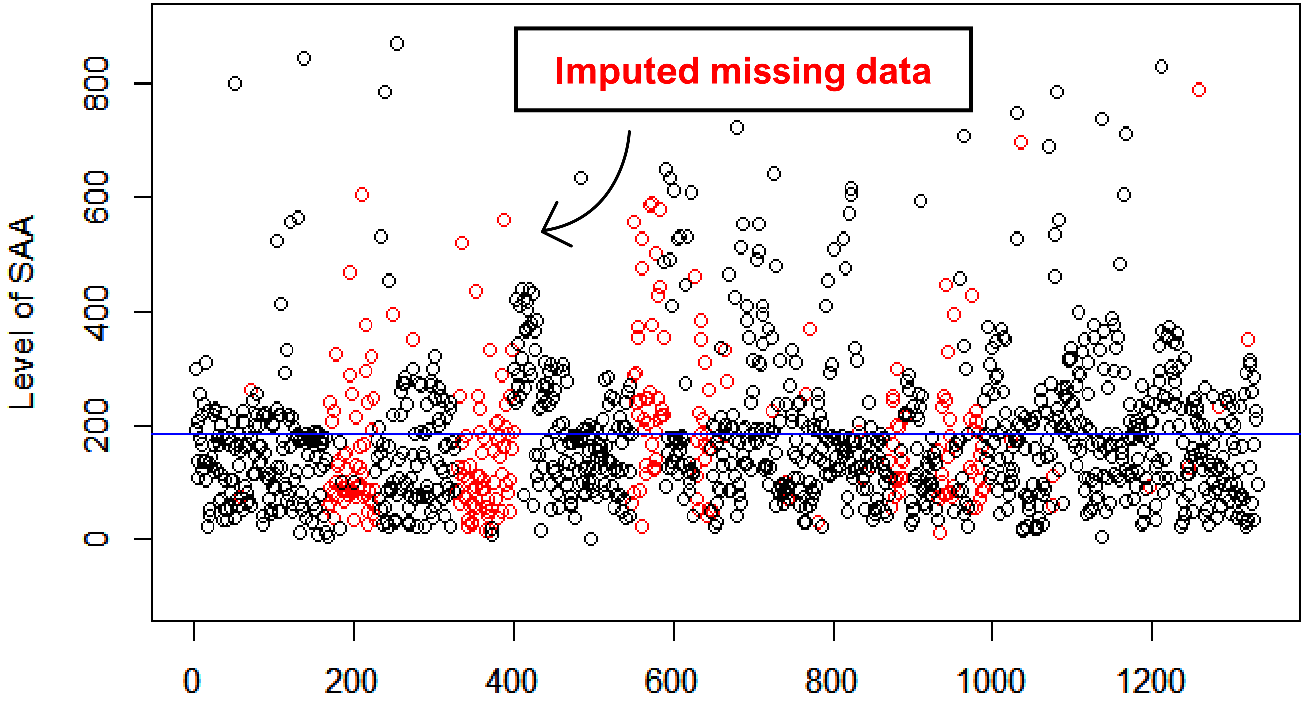
$$\boldsymbol{Ilr2} = \frac{1}{\sqrt{6}} * \log\left(\frac{GRAN * LYM}{MONO^2}\right)$$



Data processing and analysis

<u>Data processing</u>:

- Square root or logarithmic transformation of appropriate variables.
- Compositional haematological data: isometric log-ratio transformation.
- Stochastic regression imputation of missing data (MICE package).



Data processing and analysis

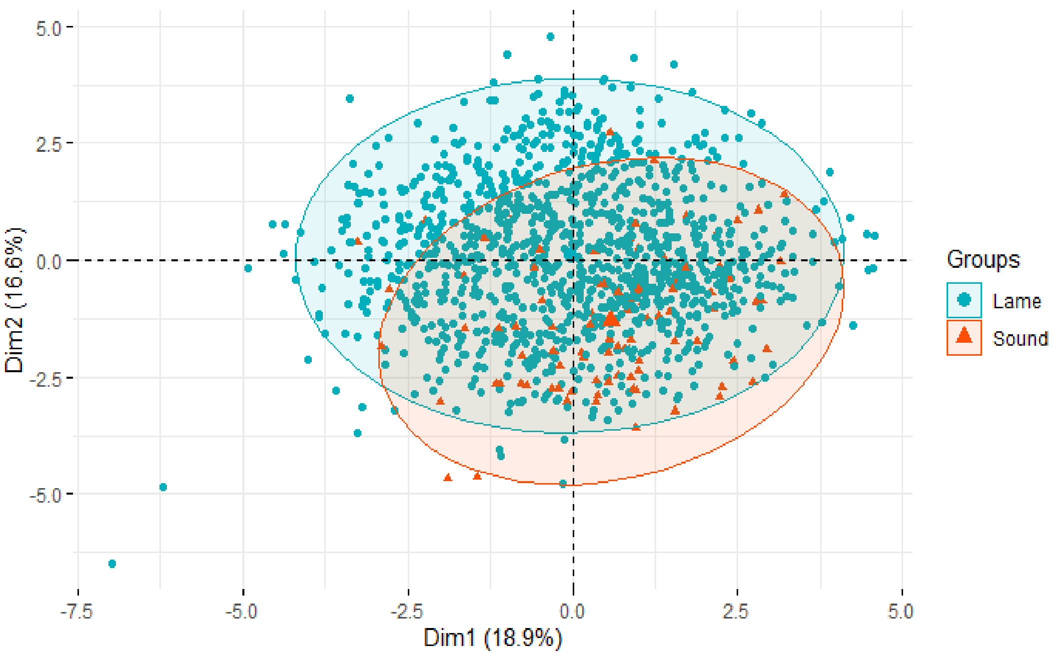
<u>Data processing:</u>

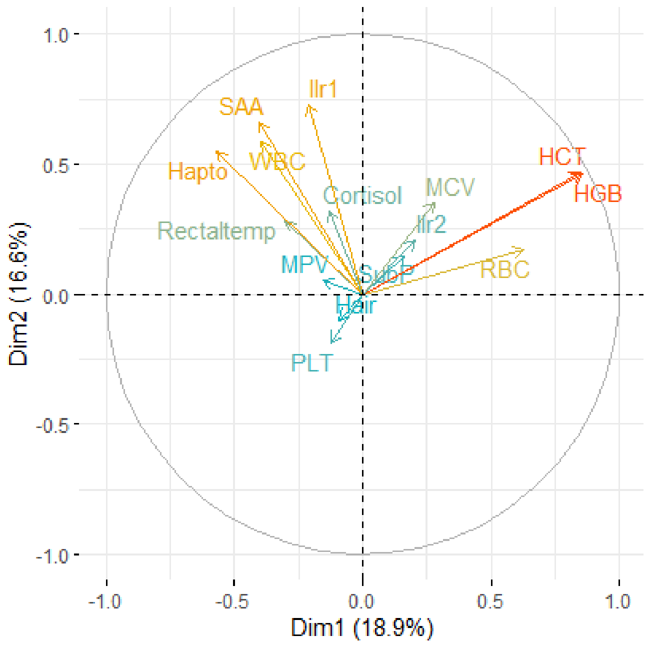
- Square root or logarithmic transformation of appropriate variables.
- Compositional haematological data: isometric log-ratio transformation.
- Stochastic regression imputation of missing data (MICE package).

<u>Data analysis:</u>

Principal component analysis (PCA)

Dimension reduction and visualisation of multivariate data.





Data processing and analysis

<u>Data processing:</u>

- Square root or logarithmic transformation of appropriate variables.
- Compositional haematological data: isometric log-ratio transformation.
- Stochastic regression imputation of missing data (MICE package).

<u>Data analysis:</u>

Principal component analysis (PCA)

Dimension reduction and visualisation of multivariate data.

Binary logistic regression

Least Absolute Shrinkage and Selection Operator **(LASSO)** method

Data processing and analysis

Data processing:
- Square root or logarithmic transformation of appropriate variables.
- Compositional haematological data: isometric log-ratio transformation.
- Stochastic regression imputation of missing data (MICE package).

Data analysis:

Principal component analysis (PCA)

Dimension reduction and visualisation of multivariate data.

Binary logistic regression

Least Absolute Shrinkage and Selection Operator (LASSO) method

Multinomial logistic regression

An extension to classical logistic regression which allows for a non-binary response variable

Data processing and analysis

<u>Data processing:</u>

- Square root or logarithmic transformation of appropriate variables.
- Compositional haematological data: isometric log-ratio transformation.
- Stochastic regression imputation of missing data (MICE package).

<u>Data analysis:</u>

Principal component analysis (PCA)

Dimension reduction and visualisation of multivariate data.

Binary logistic regression

Least Absolute Shrinkage and Selection Operator (LASSO) method

Multinomial logistic regression

An extension to classical logistic regression which allows for a non-binary response variable

Discriminant Analysis (DA)

- Linear DA
- Quadratic DA
- Distance-based DA

Data processing and analysis

Data processing:
- Square root or logarithmic transformation of appropriate variables.
- Compositional haematological data: isometric log-ratio transformation.
- Stochastic regression imputation of missing data (MICE package).

Data analysis:

Principal component analysis (PCA)

Dimension reduction and visualisation of multivariate data.

Binary logistic regression

Least absolute Shrinkage and Selection Operator (LASSO) method

TODAYS FOCUS

Multinomial logistic regression

An extension to classical logistic regression which allows for a non-binary response variable

Discriminant Analysis (DA)
- Linear DA
- Quadratic DA
- Distance-based DA

Logistic regression (1)

Logistic regression is a supervised learning method for classification. The name stems from the term "logit" which refers to the term "log odds":

$$odds = \frac{P(event)}{1 - P(event)}.$$

The response variable is a categorical variable with a binary outcome

For the probability of an event occurring $p(X) = P(Y = 1 \mid x)$ we get the conditions $p(X)\epsilon[0,1]$ and $X \in R$.

Where π is the response probability (for example the probability of a calf being lame), the response variable is expressed as the logit function $\mathbf{logit(\pi):}$

$$\mathrm{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right), \qquad \mathrm{logit}^{-1}(\pi) = \frac{e^{\pi}}{e^{\pi}+1}.$$

Logistic regression (2)

The logistic regression model can be described as :

$$\pi(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}.$$

Where $\pi(x)$ is the conditional mean, which is expressed as $\pi(x) = E(Y|x)$, where Y denotes the response variable and x denotes a value of the independent predictor variable.

The logit transformation is defined as following:

$$g(x) = ln\left[\frac{\pi(x)}{1 + \pi(x)}\right] = \beta_0 + \beta_1 x.$$

With the response variable y being distributed binomially: $y = \pi(x) + \varepsilon \sim \mathrm{Bin}(n, \pi(x))$

Logistic regression (2)

The logistic regression model can be described as :

$$\pi(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}.$$

Extension to multiple predictor variables:

$$\pi(x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{m} \beta_i x_i)}}$$

Where $\pi(x)$ is the conditional mean, which is expressed as $\pi(x) = E(Y|x)$, where Y denotes the response variable and x denotes a value of the independent predictor variable.

The logit transformation is defined as following:

$$g(x) = ln\left[\frac{\pi(x)}{1 + \pi(x)}\right] = \beta_0 + \beta_1 x.$$

With the response variable y being distributed binomially: $y = \pi(x) + \varepsilon \sim \mathrm{Bin}(n, \pi(x))$

Logistic regression (2)

The logistic regression model can be described as :

$$\pi(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}.$$

Extension to multiple predictor variables:

$$\pi(x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{m} \beta_i x_i)}}$$

Where $\pi(x)$ is the conditional mean, which is expressed as $\pi(x) = E(Y|x)$, where Y denotes the response variable and x denotes a value of the independent predictor variable.

The logit transformation is defined as following:

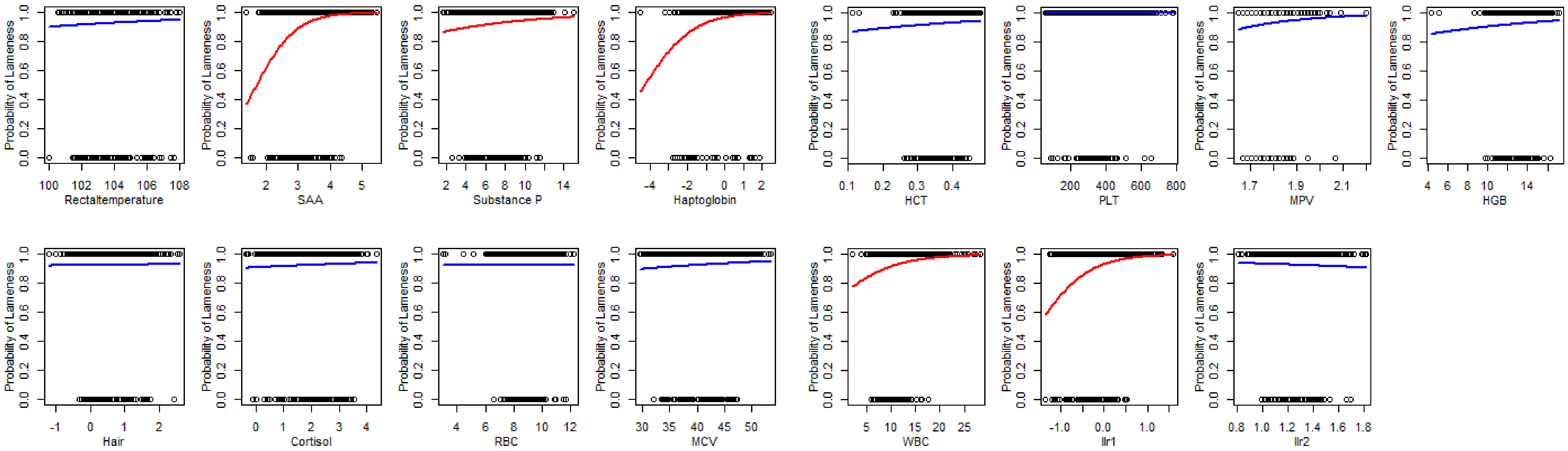
$$g(x) = ln\left[\frac{\pi(x)}{1 + \pi(x)}\right] = \beta_0 + \beta_1 x.$$

Fitting of $\boldsymbol{\beta}$: maximum likelihood function

With the response variable y being distributed binomially: $y = \pi(x) + \varepsilon \sim \mathrm{Bin}(n, \pi(x))$

Logistic regression model *sound/lame*

One predictor variable at a time



⟶ SAA, substance P, haptoglobin, white blood cell count and the granulocyte to leucocyte ratio are all by themselves significant predictors

Logistic regression models for sound/lame

	OR	2.5 %	97.5 %	pValue	Significance
(Intercept)	9.6e-07	6.4e-20	3.8e+06	0.57	
Rectaltemp	0.95	0.76	1.19	0.65	
SAA	2.08	1.25	3.52	0.01	**
SubP	1.12	0.98	1.28	0.09	.
Hapto	1.61	1.25	2.12	4.1e-05	***
Hair	1.08	0.69	1.69	0.75	
Cortisol	0.86	0.63	1.18	0.35	
RBC	4.82	0.56	51.52	0.17	
MCV	1.41	0.88	2.38	0.18	
HCT	5.8e-14	4.1e-42	2.7e+12	0.34	
PLT	1.00	1.00	1.00	0.25	
MPV	18.53	0.26	1655.69	0.19	
HGB	0.87	0.42	1.80	0.71	
WBC	0.96	0.88	1.05	0.40	
Ilr1	4.12	2.25	7.65	5.4e-06	***
Ilr2	4.8e-01	8.7e-02	2.7e+00	0.40	

$$L0 \sim SAA + SubP + Hapto + Ilr1$$

	OR	2.5 %	97.5 %	pValue	Significance
(Intercept)	1.54	0.25	9.69	0.64	
SAA	1.89	1.15	3.13	0.01	*
SubP	1.09	0.98	1.21	0.10	.
Hapto	1.65	1.29	2.14	8.9e-05	***
Ilr1	3.15	1.87	5.34	1.8e-05	***

Logistic regression models for sound/lame

	OR	2.5 %	97.5 %	pValue	Significance
(Intercept)	9.6e-07	6.4e-20	3.8e+06	0.57	
Rectaltemp	0.95	0.76	1.19	0.65	
SAA	2.08	1.25	3.52	0.01	**
SubP	1.12	0.98	1.28	0.09	.
Hapto	1.61	1.25	2.12	4.1e-05	***
Hair	1.08	0.69	1.69	0.75	
Cortisol	0.86	0.63	1.18	0.35	
RBC	4.82	0.56	51.52	0.17	
MCV	1.41	0.88	2.38	0.18	
HCT	5.8e-14	4.1e-42	2.7e+12	0.34	
PLT	1.00	1.00	1.00	0.25	
MPV	18.53	0.26	1655.69	0.19	
HGB	0.87	0.42	1.80	0.71	
WBC	0.96	0.88	1.05	0.40	
Ilr1	4.12	2.25	7.65	5.4e-06	***
Ilr2	4.8e-01	8.7e-02	2.7e+00	0.40	

$$L0 \sim SAA + SubP + Hapto + Ilr1$$

	OR	2.5 %	97.5 %	pValue	Significance
(Intercept)	1.54	0.25	9.69	0.64	
SAA	1.89	1.15	3.13	0.01	*
SubP	1.09	0.98	1.21	0.10	.
Hapto	1.65	1.29	2.14	8.9e-05	***
Ilr1	3.15	1.87	5.34	1.8e-05	***

Odds ratio (OR)

$$\mathrm{OR} = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} = e^{\beta_1}$$

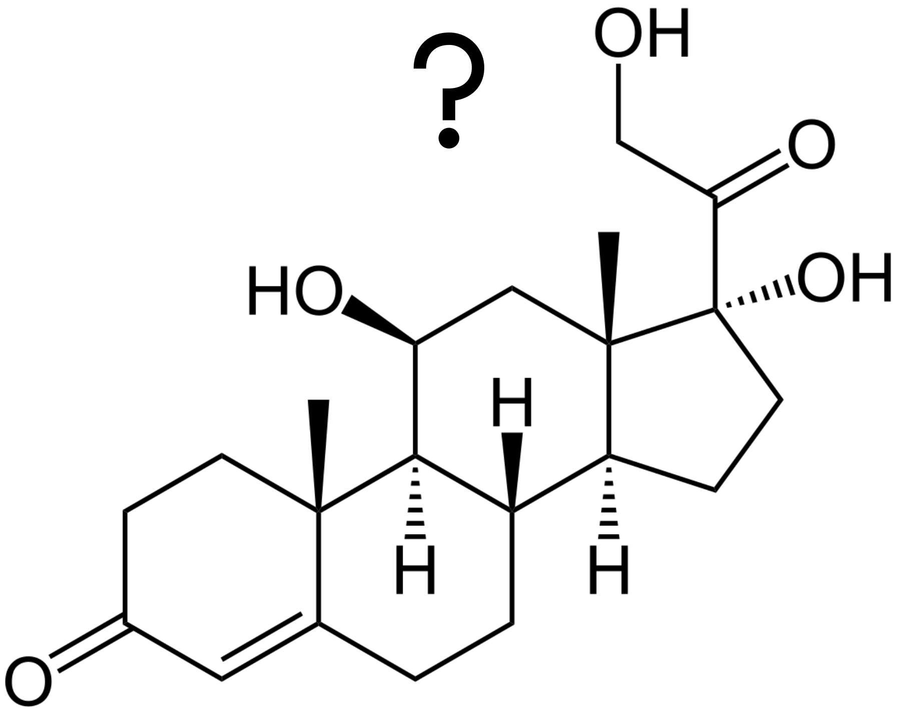
Indicators of calf lameness

- **Haptoglobin** and **serum amyloid A (SAA)** showed significantly higher values in calves affected with lameness than in healthy individuals in this study.

- Neither **blood** nor **hair cortisol** proved to be significant predictors for lameness in calves.

  ⟶ **Substance P** however did.

- Ilr1 representing the **ratio of granulocytes to lymphocytes** also resulted to be a significant predictor for lameness in calves.

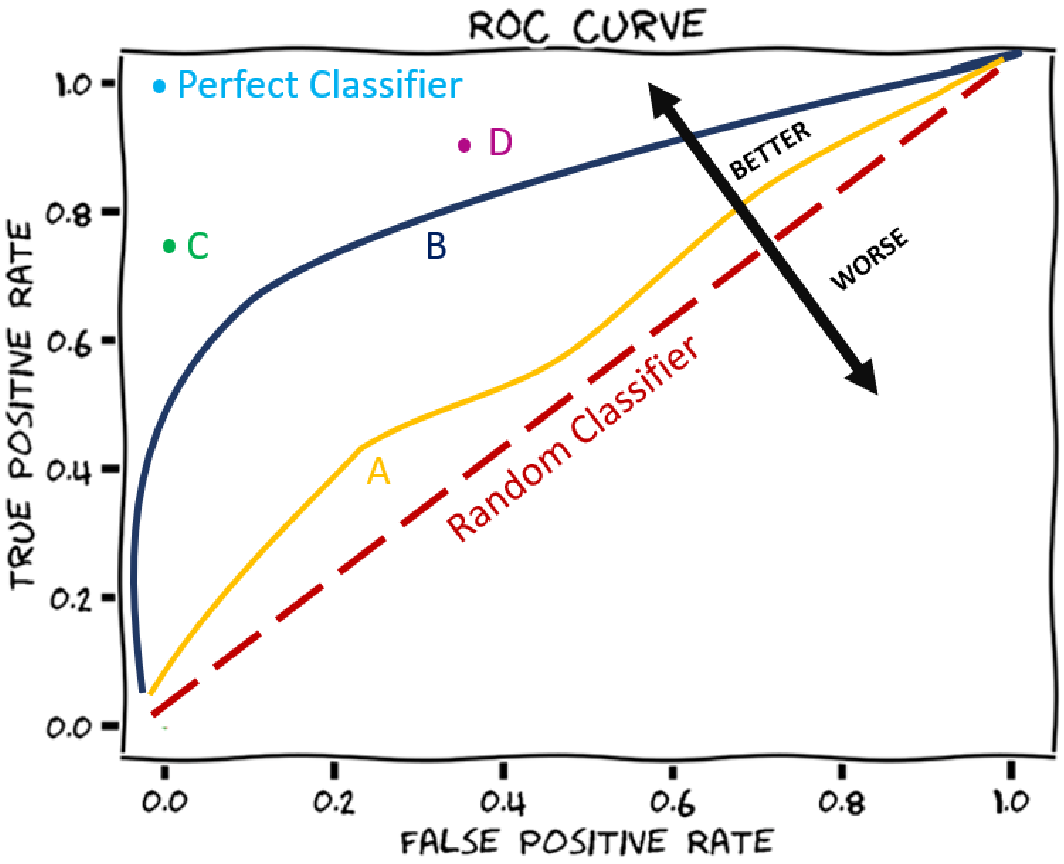


Validation of binary prediction models

Receiver operating characteristics (ROC) curves show a relative trade-off between benefits (true positives) and costs (false positives).

$$TPR = \frac{true\ positives}{number\ of\ total\ positives}$$

$$FPR = \frac{false\ positives}{number\ of\ total\ negatives}$$

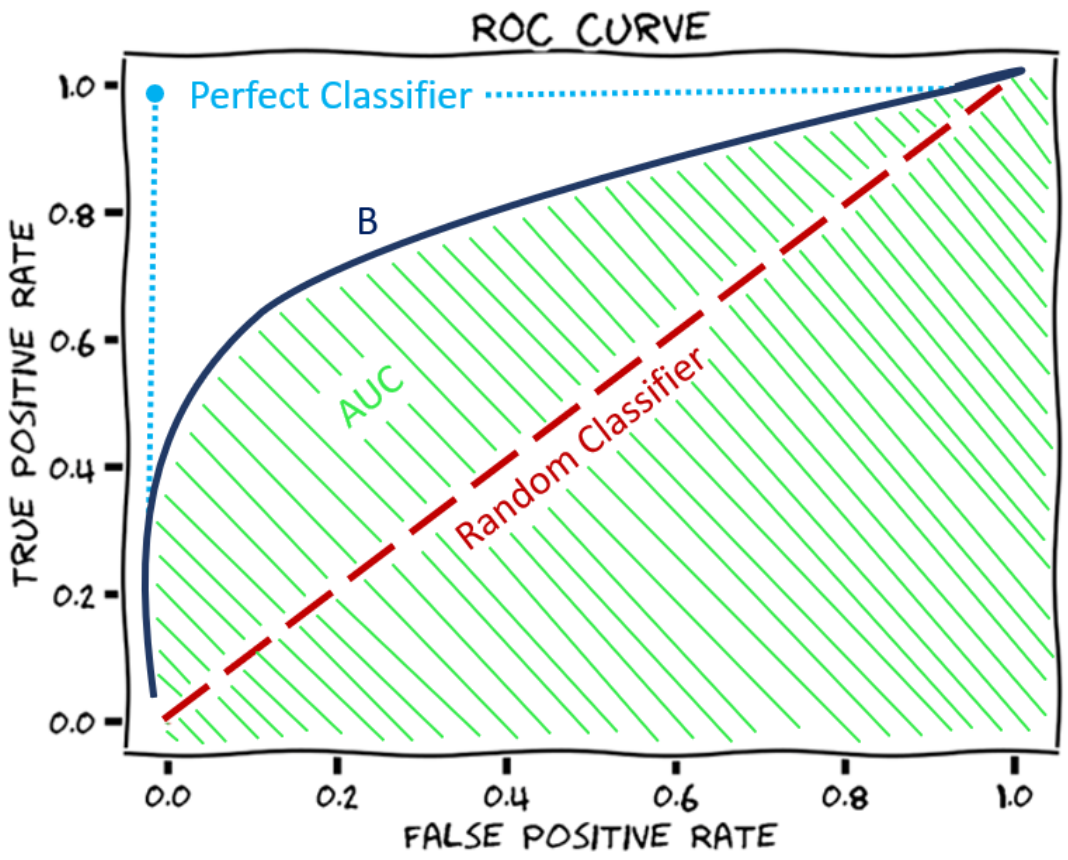


modified after Wikipedia on ROC space
https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Validation of binary prediction models

Receiver operating characteristics (ROC) curves show a relative trade-off between benefits (true positives) and costs (false positives).

Area Under Curve (AUC) score: standard measure of accuracy for assessing the performance of binary predictive models.



modified after Wikipedia on ROC space
https://en.wikipedia.org/wiki/Receiver_operating_characteristic

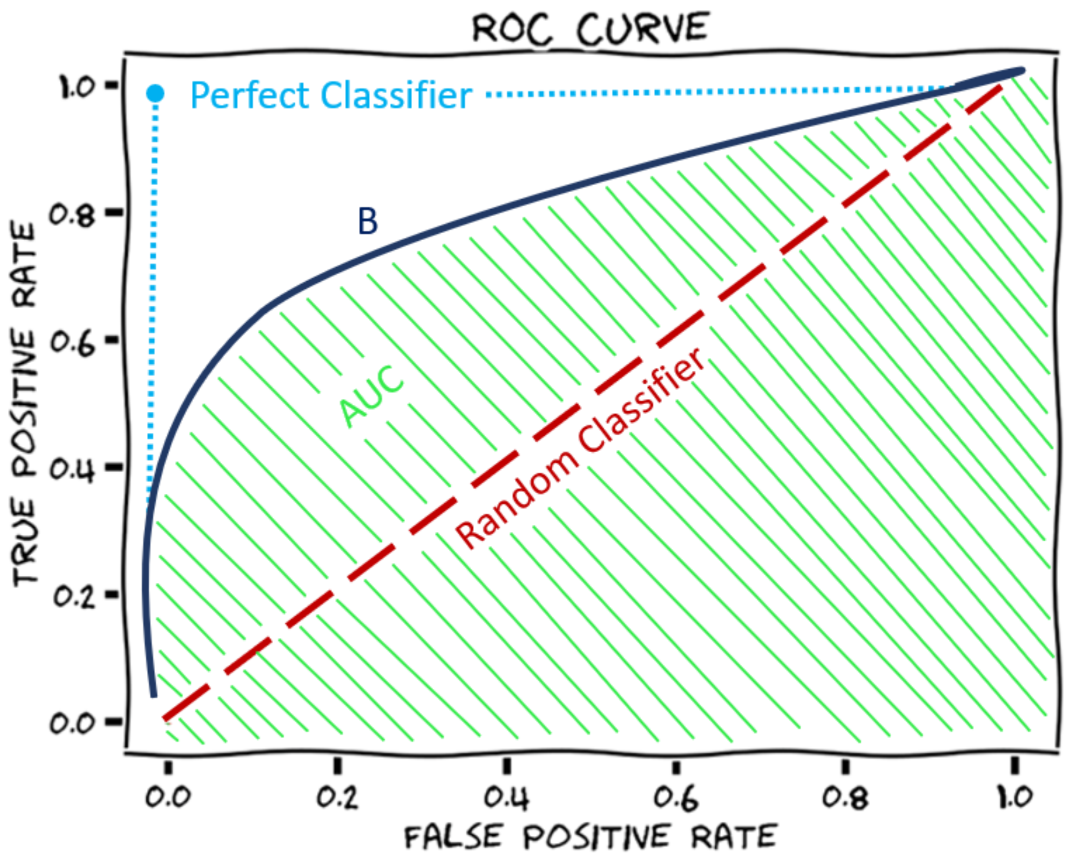
Validation of binary prediction models

Receiver operating characteristics (ROC) curves show a relative trade-off between benefits (true positives) and costs (false positives).

Area Under Curve (AUC) score: standard measure of accuracy for assessing the performance of binary predictive models.

Gini coefficient is used for comparing the quality of different models and prediction power.

$$Gini = (2\,AUC\ - 1) * 100$$



modified after Wikipedia on ROC space
https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Gini coefficient vs. Accuracy

Gini coefficient is used for comparing the quality of different models and prediction power.

$$Gini = (2\,AUC - 1) * 100$$

Accuracy as calculated from the confusion matrix:

$$Accuracy = \frac{true\ positives + true\ negatives}{total\ observations} * 100$$

Problematic in unbalanced class situations

	All Observations *Lame/Sound = 12.5/1* *n = 1271*			Test Set *Lame/Sound = 1/1* *n = 186*		
Confusion Matrix		0	1		0	1
	P0	9	9	P0	75	30
	P1	84	1194	P1	18	63
Accuracy	**92.68%**			**74.19%**		
AUC	0.81			0.80		
Gini	62.63			60.04		

Note: rows correspond to the prediction and columns to the actual state of the cattle.

Logistic regression models for every lesion type

	Sound	Foot Rot	Dermatitis	P3 Necrosis	Proximal Limb Issue	Injury	Joint Infection	Other
(Intercept)	0.43	0.05 ***	0.02 ***	1.1e-26 ***	1.9e+17 **	3.3e+12 *	0.004 ***	7.8e+12 **
Rectaltemp				1.80***	0.77*			
SAA	0.49**	1.75**	0.70*				1.36	
SubP		1.15***		0.66***				
Hapto	0.62***	1.42***	0.66***		1.20*			0.51**
Cortisol		0.81*						
RBC				9.82***		0.03*		
MCV						0.46*		
HCT		2.4e-9***				19.4e+37*		
PLT							1.004***	
MPV					0.005*			1.1e-08**
HGB		1.64**	1.58***	0.16***	0.58***			0.82
WBC		1.22***	0.85***					
Ilr1	0.32***		3.43***					
Ilr2								
Hair		0.63***			1.79*			
AUC	0.81	0.77	0.79	0.95	0.75	0.57	0.65	0.84
Gini	62.20	54.38	57.45	90.98	51.95	14.62	29.74	68.61

Logistic regression models for every lesion type

	Sound	Foot Rot	Dermatitis	P3 Necrosis	Proximal Limb Issue	Injury	Joint Infection	Other
(Intercept)	0.43	0.05 ***	0.02 ***	1.1e-26 ***	1.9e+17 **	3.3e+12 *	0.004 ***	7.8e+12 **
Rectaltemp				1.80***	0.77*			
SAA	0.49**	1.75**	0.70*				1.36	
SubP		1.15***		0.66***				
Hapto	0.62***	1.42***	0.66***		1.20*			0.51**
Cortisol		0.81*						
RBC				9.82***		0.03*		
MCV						0.46*		
HCT		2.4e-9***				19.4e+37*		
PLT							1.004***	
MPV					0.005*			1.1e-08**
HGB		1.64**	1.58***	0.16***	0.58***			0.82
WBC		1.22***	0.85***					
Ilr1	0.32***		3.43***					
Ilr2								
Hair		0.63***			1.79*			
AUC	0.81	0.77	0.79	0.95	0.75	0.57	0.65	0.84
Gini	62.20	54.38	57.45	90.98	51.95	14.62	29.74	68.61

LASSO logistic regression (1)

LASSO (Least Absolute Shrinkage and Selection Operator) adds a penalty term to the log likelihood function ℓ used to find logistic regression coefficients $\boldsymbol{\beta}$.

The penalty term is $\boxed{\lambda \sum |\boldsymbol{\beta}_j|}$. The quantity to be minimised in the two cases is thus:

$$\underbrace{\boxed{\ell_2^L(\beta) = \sum_{i=1}^{n} \left[y_i \beta x_i - \ln\left(1 + e^{\beta x_i}\right) \right]}}_{\text{Maximum likelihood function}} - \boxed{\lambda \sum_{j=1}^{m} |\boldsymbol{\beta}_j|.}$$

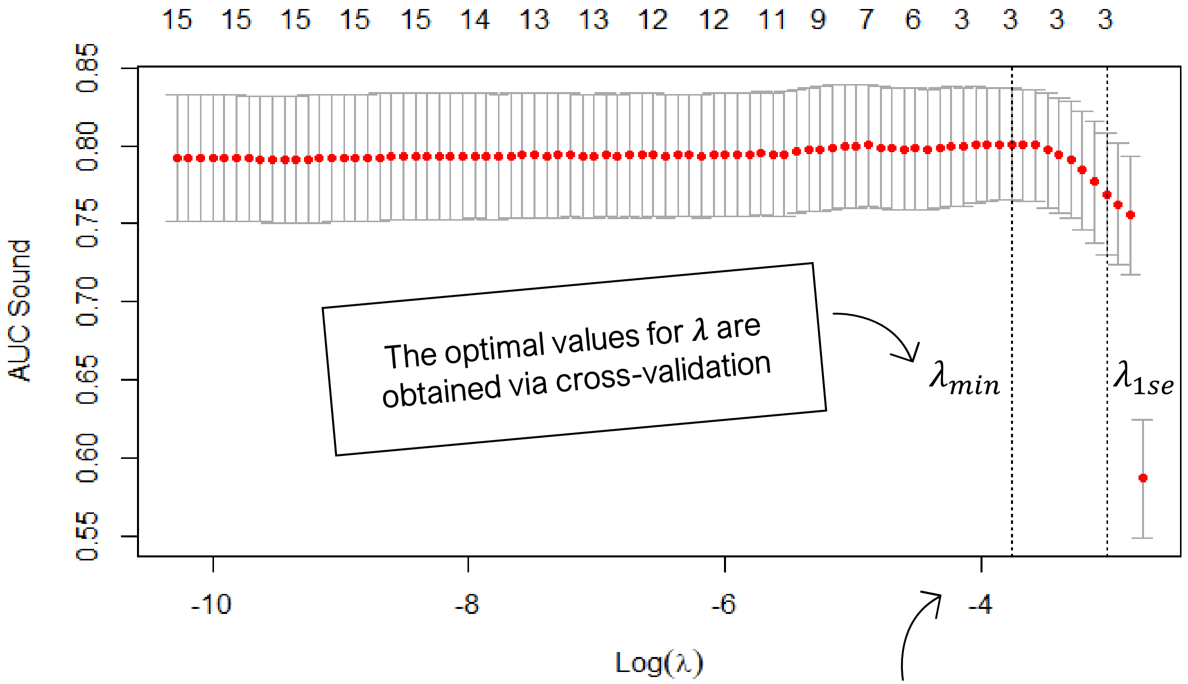
Where λ is a free parameter, selected in such a way that the resulting model minimises the out of sample error. Typically, the optimal value of λ is found using grid search with cross-validation

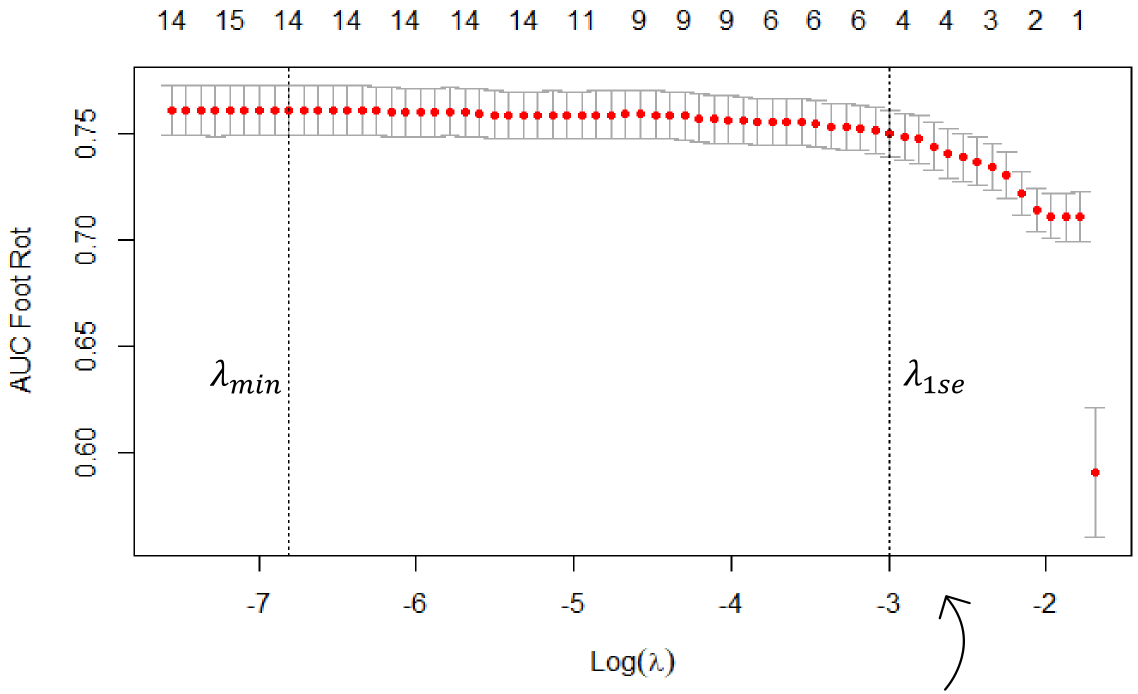
LASSO logistic regression (2)

$\longrightarrow$ "cv.glmnet" function from R package "glmnet".
(cv = cross-validation with 5 folds)

This function offers two options for the penalty term $\lambda \sum|\boldsymbol{\beta}_j|$.

LASSO avoids model overfitting by the introduction of the penalty term





λ_{min} minimizes out-of-sample loss in cross validation

λ_{1se} is the largest value within 1 standard error of λ_{min}

LASSO logistic regression models for every lesion type

	Sound	Foot Rot	Dermatitis	P3 Necrosis	Proximal Limb Issue	Injury	Joint Infection	Other
(Intercept)	-0.84	-3.14	1.12	-62.63	40.24	-2.05	17.62	25.68
Rectaltemp		0.03	-0.07	0.58	-0.29		-0.18	-0.20
SAA	-0.58	0.47	-0.42	0.08	-0.12	0.11	0.33	0.44
SubP		0.14	0.01	-0.36	0.04	-0.11	-0.08	-0.16
Hapto	-0.26	0.35	-0.40	0.16	0.29			-0.94
Cortisol		-0.21	0.02	0.14	0.22			0.42
RBC		-0.18	-0.42	2.76	0.74		0.04	2.71
MCV				0.18	0.08		-0.06	0.68
HCT		-14.65			20.13			41.77
PLT								0.01
MPV		-0.61	1.54	-1.67	-5.95		-1.01	-21.42
HGB		0.43	0.65	-2.21	-1.63			-3.35
WBC		0.12	-0.13	-0.06	-0.08			
Ilr1	-0.73	0.02	1.05	-0.56	0.03			-0.20
Ilr2		-0.73	0.33	-0.70	1.59		0.35	
Hair		-0.38	0.28	0.45	0.55			-0.39
Gini (before)	62.20	54.38	57.45	90.98	51.95	14.62	29.74	68.61
Gini LASSO	62.69	55.11	60.07	92.13	**58.22**	**36.05**	**39.61**	**86.39**

LASSO logistic regression models for every lesion type

	Sound	Foot Rot	Dermatitis	P3 Necrosis	Proximal Limb Issue	Injury	Joint Infection	Other
(Intercept)	-0.84	-3.14	1.12	-62.63	40.24	-2.05	17.62	25.68
Rectaltemp		0.03	-0.07	0.58	-0.29		-0.18	-0.20
SAA	-0.58	0.47	-0.42	0.08	-0.12	0.11	0.33	0.44
SubP		0.14	0.01	-0.36	0.04	-0.11	-0.08	-0.16
Hapto	-0.26	0.35	-0.40	[illegible]	[illegible]	[illegible]		-0.94
Cortisol		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		0.42
RBC		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.04	2.71
MCV		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.06	0.68
HCT		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		41.77
PLT		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		0.01
MPV		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-1.01	-21.42
HGB		[illegible]	[illegible]	-2.21	-1.63			-3.35
WBC		0.12	-0.13	-0.06	-0.08			
Ilr1	-0.73	0.02	1.05	-0.56	0.03			-0.20
Ilr2		-0.73	0.33	-0.70	1.59		0.35	
Hair		-0.38	0.28	0.45	0.55			-0.39
Gini (before)	62.20	54.38	57.45	90.98	51.95	14.62	29.74	68.61
Gini LASSO	62.69	55.11	60.07	92.13	**58.22**	**36.05**	**39.61**	**86.39**

By employing Lasso models on a test set, we were able to predict the right lesion (out of 8 options) with an **accuracy of 61%**.

Multinomial logistic regression

Multicategory logit models simultaneously use all pairs of categories by specifying the odds of outcome in one category instead of another.

Y_i is a categorical and polytomous response variable with multiple classes k. One class is set as the baseline category Y_0. Furthermore we use m explanatory variables (or predictors):

$$\log\left(\frac{\pi_i^{(Y_i)}}{\pi_i^{(Y_0)}}\right) = \alpha^{(Y_i)} + \beta_1^{(Y_i)} X_{1i} + \cdots + \beta_m^{(Y_i)} X_{mi}$$

Where $r = 1, \dots, k$, the multinomial logistic regression model can be described :

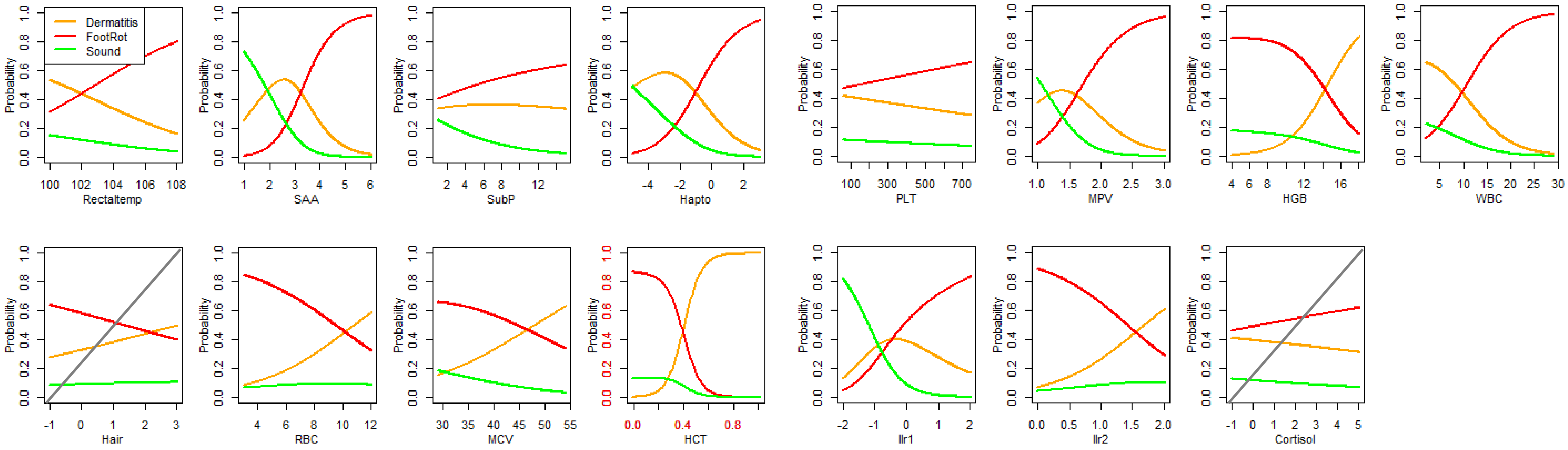
$$\pi_{ir} = \mathrm{P}(Y_i = r) = \frac{e^{\left(\alpha_{r0} + \boxed{\mathbf{x}_i^{\top}\boldsymbol{\beta}_r}\right)}}{\sum_{s=1}^{k} e^{\left(\alpha_{s0} + \boxed{\mathbf{x}_i^{\top}\boldsymbol{\beta}_s}\right)}}$$

$$\boxed{\mathbf{x}_i^{\top}\boldsymbol{\beta}_r = \beta_{r0} + \beta_{r1}x_{i1} + \beta_{r2}x_{i2} + \cdots + \beta_{rm}x_{im}}$$

$$\boxed{\mathbf{x}_i^{\top}\boldsymbol{\beta}_s = \beta_{s0} + \beta_{s1}x_{i1} + \beta_{s2}x_{i2} + \cdots + \beta_{sm}x_{im}}$$

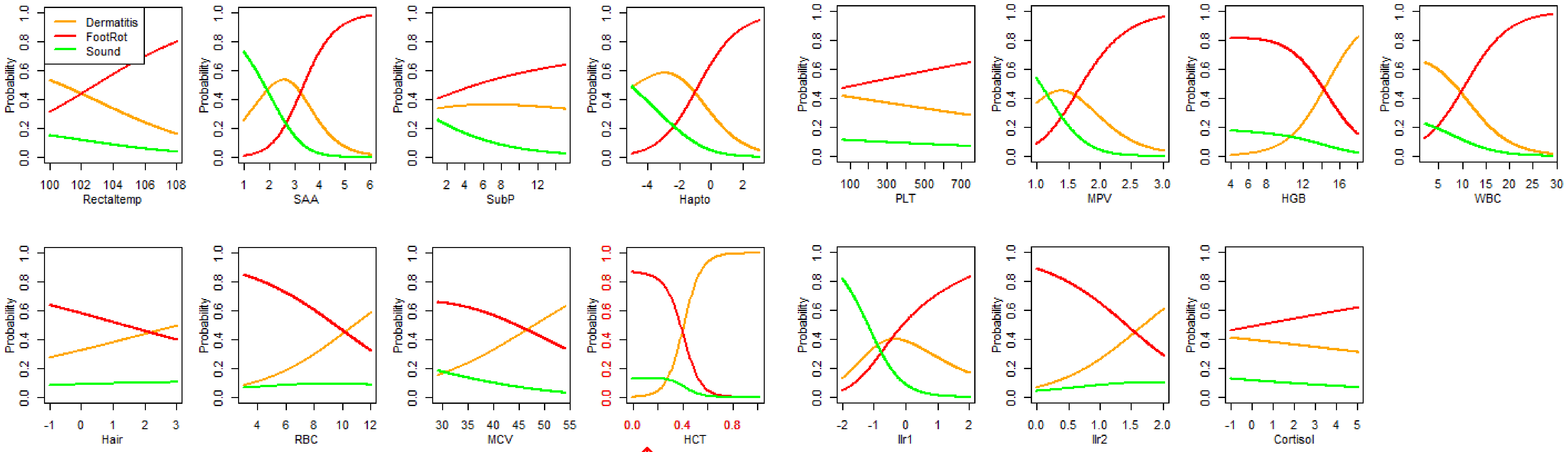
Multinomial logistic regression

Visualisation of predicted probability



Multinomial logistic regression

Visualisation of predicted probability

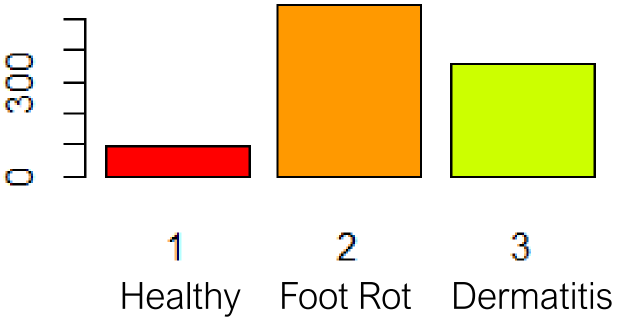


Haematocrit (HCT) acts as a very good distinguisher between foot rot and dermatitis

Multinomial logistic regression vs. discriminant analysis

Both model types can be used to predict with non-binary response variables.

We compare their accuracy using a subset (n = 992) of the two most common classes and the healthy animals:



Train / test partition with 60% train and 40% test set.

	Small data set / n = 992				
		1	2	3	
Multinomial logistic regression	P1	26	5	7	Accuracy: 70.9% / **test:** 73%
	P2	22	446	120	
	P3	45	90	231	
		1	2	3	
Linear discriminant analysis	P1	25	8	14	Accuracy: 69% / **test**: 72%
	P2	18	427	110	
	P3	50	106	234	
		1	2	3	
Quadratic discriminant analysis	P1	51	13	15	Accuracy: 75% / **test:** 69%
	P2	13	421	75	
	P3	29	107	268	

Note: rows correspond to the prediction and columns to the actual state of the cattle.

Future extensions and improvement

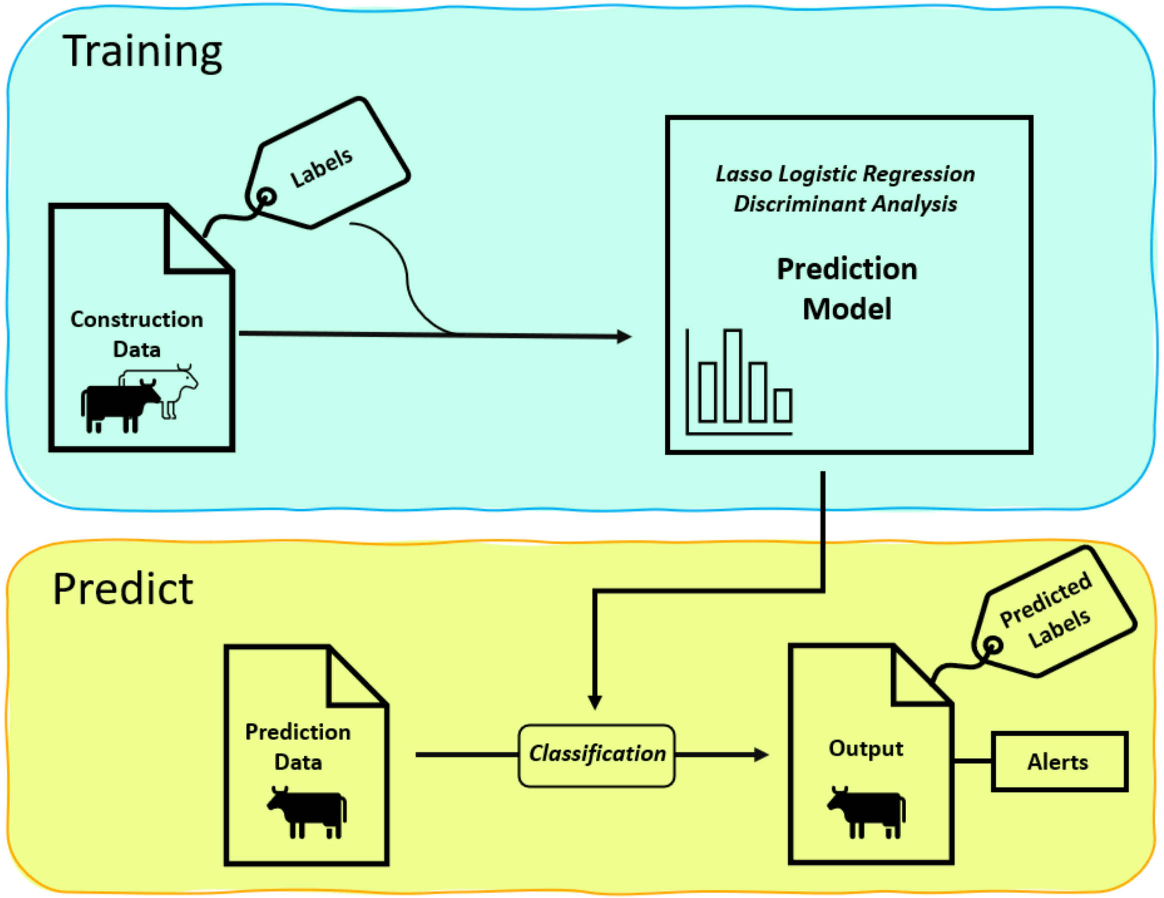
Main problems: unbalanced-class situation and misclassification between foot rot and dermatitis.

Extension of AUC score to non-binary models: **generalised AUC score.**

Employment of alert counter: Binary variable representing variables that encode risk factors but are activated only on a small portion of our samples (decision tree with one level).

⟶ **Random forest**

+ gradient boosting



https://bestlifeonline.com/cow-jokes/