ADVANCED ANALYTICAL AND MACHINE LEARNING METHODS FOR ANALYSIS OF SELECTION AND PREDICTION OF MORTALITY IN COMMERCIAL SWINE

A Dissertation

Presented to the Faculty of the Graduate School

at the University of Missouri – Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

CALEB J. GROHMANN

Dr. Jared E. Decker, Dissertation Supervisor

JULY 2024

APPROVAL PAGE

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

ADVANCED ANALYTICAL AND MACHINE LEARNING METHODS FOR ANALYSIS OF SELECTION AND PREDICTION OF MORTALITY IN COMMERCIAL SWINE

Presented by Caleb J. Grohmann, a candidate for the degree of Doctor of Philosophy, and hereby certify that in their opinion it is worthy of acceptance.

Dr. Jared E Decker, Animal Science, UMC

Dr. Robert D. Schnabel, Animal Science, UMC

Dr. Timothy J. Safranski, Animal Science, UMC

Dr. Elizabeth King, Biological Science, UMC

Dr. Guilherme DeSouza, Computer Science, UMC

DEDICATION

Ever since I could walk, my dad, Freddie Grohmann, has shown me how to be a pig farmer. Without his guidance and knowledge surrounding the entire swine industry, I would not have been able to pursue a successful career in commercial pig production. Because of him and growing up on Cedar Ridge Farms in southern Illinois, I gained invaluable skills and an understanding of the swine industry that I utilize every day in my current work and career. Most importantly, he has been a perfect role model of how to be a professional in agriculture, as well as an outstanding father. I dedicate this dissertation to my father, Frederick Grohmann. I can only hope to have the same impact on the international swine industry as you have!

ACKNOWLEDGEMENTS

Where do I begin? There are so many people that have been integral to my entire academic career, and it has truly been a team effort. First, I would like to thank my entire family, Freddie, Leslie, Frederick, Lexi, and Austin Grohmann and my grandparents Jerry and Lois Moll. Your support throughout the years in good times and in bad has been fully appreciated, and I am indebted to all of you for sticking with me through this lifelong process. Next, I would like to thank my best friend Isabella Sellmer Ramos, whose been a perfect partner for the past two years as well as an amazing scientific mind to always keep me on the right track. I am excited for this next chapter of our life and for all the amazing things you will do with your career as well.

My academic career has spanned many different schools. I would like to thank the late Dan Flowers, my high school math teacher, for providing a basis in which I could grow my statistics knowledge. In addition, Coach David Gillingham taught me valuable life, leadership, and team-building skills during my time on the Red Bud High School basketball team. At the University of Missouri, I am indebted to Dr. Tim Safranski, who was key to me attending the University of Missouri and was and is a great mentor and knowledgeable about all things pigs.

Next, Dr. Michael Ellis, my Master's advisor at the University of Illinois, provided a foundational understanding of commercial pig production and applied, real-world research. Finally, Dr. Jared Decker has been an outstanding PhD mentor. Jared has allowed me to work independently and has promoted any idea I may have had over the past 4 years. Also, thank you to the rest of my committee, Dr. Robert Schnabel, Dr.

Guilherme DeSouza, and Dr. Elizabeth King for their guidance with my PhD dissertation and research.

There have been several friends that have helped shape my life as well. To the roommates of the “House of Learning Doctors”, Drs. Bryce and Katy McDonald, Dr. Rachael Bonacker, and Dr. Sara Schroer, without you guys, my time as an undergraduate and PhD student would have not been the same. I was lucky to share so much time with like-minded individuals, and the good times we have had always kept me persevering. Next, without my most recent roommate Timmy Rackers, I would not have had a place to stay or as many laughs to share as I have over the past two years. Finally, Matt Haas has been a great friend and co-owner of Lakeside Farms, our Boer goat operation. Without your assistance, our farm could not have survived while I was in school, and I am looking forward to its future growth.

None of this work would have been possible without the team throughout the years at The Maschhoff’s, Dr. Caleb Shull, Dr. Bradley Wolter, Dr. Clint Schwab, Dr. Beau Peterson, Dr. Alysta Sewell, Kristy Johnson, and Randy Bowman. These men and woman first introduced me to The Maschhoff’s in high school, and have helped stimulate my career in many ways through their mentorship in my Master’s program and several different internships. In addition, the funding and access to pigs and data provided by The Maschhoff’s and the Foundation for Food and Agricultural Research was very important and much appreciated over my graduate school career. I thank you all for your past and continued support!

Last but not least, I would like to thank all fellow lab members during my Master’s and PhD programs, Dr. Jenny Morris, Heath Harper, Andres Tolosa, Dr.

Katherine Vande Pol, Ovidio Bautista, Dr. Harly Durbin, Dr. Troy Rowan, AJ Knowles, John Miraszek, and Dr. Jenna Kallenberg. I would not have made it through without their collaboration and friendship!

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY FIGURES

Figures **Page**

LIST OF SUPPLEMENTARY TABLES

TABLE OF CONTENTS

ABSTRACT

The age of information and the Internet of Things (IoT) has brought forth many exciting opportunities for farmers and researchers in commercial swine production. The amount of data available across all sectors of the industry is rapidly increasing, which requires innovative methods to store, analyze, and derive insights to positively impact producer economic sustainability. In chapter one of this dissertation, we propose a framework for shifting from reactive to proactive livestock management, which is assisted by technologies in artificial intelligence. Further, in chapter two, a novel method known as generation proxy selection mapping (**GPSM**) was utilized to identify single nucleotide polymorphisms in a commercial pig population that are undergoing significant changes in allele frequency over short time scales (i.e., four to ten years). In chapter three, we developed an algorithm to identify periods of episodic mortality in commercial wean-to-finish pig cohorts, which reveal sequences of days in which mortality is acutely increased relative to population baseline. Lastly, in chapter four, we evaluated various machine learning models to forecast episodic and sporadic mortality in growing pigs. Results from this work can promote evidence-based, data-driven decision making in commercial pig production in real-time.

CHAPTER 1. FROM REACTIVE TO PROACTIVE: IMPACT OF ARTIFICIAL INTELLIGENCE ON MANAGEMENT AND SELECTION OF LIVESTOCK

Caleb J. Grohmann[†]* and Jared E. Decker [†‡§]*

[†]Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA

[‡]Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA

[§]Genetics Area Program, University of Missouri, Columbia, MO 65211, USA

*Corresponding authors

IMPLICATIONS

- Collection of automated measurements using sensors, cameras, and production data enables a shift from reactive to proactive management of livestock systems.
- Machine learning and artificial intelligence can incorporate diverse data streams into a single model that yields a prediction for key performance indicators of animal health, well-being, and productivity.
- Farmers, ranchers, and managers can utilize real-time daily forecasts from artificial intelligence systems to devise intervention plans to embrace proactive management.
- Phenotypes measured by these sensors or predicted by artificial intelligence will allow selection for drivers of sustainability, rather than low-information indicator traits.
- Reducing time to intervention will mitigate the effect of health issues on the sustainability of farms and ranches.

INTRODUCTION

As innovation and advances in technology have exponentially increased over the past several years, the amount of raw data that has been collected, stored, and analyzed has risen at a commensurate rate (Morota et al., 2018). Few, if any, industries are immune to these advances, as technology and innovation are critical for the economic viability and sustainability of most businesses. Agriculture, especially in crop production, has been an early adopter of precision management techniques, which are usually powered by artificial intelligence (**AI**) tools to assist crop management and decision support. However, in the livestock sector, the adoption of AI tools has been considerably less widespread.

There are several technologies that utilize AI in livestock production to quantify a current outcome in a production system. Examples of these technologies are cameras that estimate the current body weight of pigs (Vranken and Berckmans, 2017; Morota et al., 2018), microphones to assess cough incidence in wean-to-finish pig barns (Vranken and Berckmans, 2017), or accelerometers that track overall animal activity levels (Vázquez Diosdado et al., 2015). Within these sensors and cameras, machine learning models, typically based on artificial neural networks, take in signals (e.g., audio, video, etc.) captured by the device, process the input, and output a prediction for each respective outcome (e.g., body weight, cough incidence, activity level, etc.). Livestock producers then use these predictions to decide whether to deviate from standard management protocols. While there is little doubt that inclusion of AI-assisted support tools in livestock production systems will increase in future years, the sector is currently in a "early-adopter" phase, where only the most innovative farms and ranches have

implemented monitoring systems based on artificial intelligence (Vranken and Berckmans, 2017). Unfortunately, most of these technologies only currently provide a snapshot of the health and productivity of the animals at a given point in time, as opposed to a forecast of future metrics that would drive profitable decisions.

Historically, livestock producers have managed farms and ranches reactively as opposed to proactively. Artificial intelligence can promote a behavioral change towards proactive livestock production. By combining sensors, cameras, and production data in a cohesive infrastructure, livestock producers will have predictions to anticipate detrimental factors that influence animal health, well-being, and overall economic sustainability.

REACTIVE LIVESTOCK PRODUCTION

In worst case scenarios, livestock managers have not evaluated the efficiency or profitability of their operation and have only reactively acted when issues arose. Others have used "close-out" reports to retrospectively evaluate the performance of their production systems. Regardless of species, these reports only consider historical data for the previous sold lot, finishing cohort, birth season, or lactation period. These reports, while still important, enable a reactive decision-making process, where changes to management plans are enacted based on previous groups of animals that are generally not optimal for future groups of animals. Unfortunately, in many genetic evaluations, traits have been chosen for predictions based on ease of measurement, rather than a strong relationship to sustainability outcomes. The logical next step is to track production metrics daily as the animals are in the growing or lactation phase. However, this is also generally an inadequate approach, as daily time series from multiple health and

productivity indicators are complex and detrimental patterns are hard to recognize by manual inspection. Sensors can help farm staffs monitor barns or individual animals more closely. But, as the number of sensors used increases, the decision to change management strategies or intervene during adverse health events becomes more complex, especially if sensors provide conflicting assessments. Further, outcomes may result from interactions of the measurements from multiple sensors. Early warning systems using sensors have been previously proposed and evaluated for individual indicators (Vranken and Berckmans, 2017). However, for a more complex variable such as mortality, a network of diverse data streams will be necessary to maximize forecasting accuracy and optimize intervention plans.

The few complications stated above, amongst many others, have contributed to the slow adoption of artificial intelligence in the animal sector, which perpetuates reactive livestock production. The next frontier is proactive livestock production, and artificial intelligence is integral to this required behavioral shift.

AI-ASSISTED PROACTIVE LIVESTOCK PRODUCTION

Sustainability, defined as profitability, social responsibility, and environmental impact, is key for livestock producers, and animal health and well-being are key to sustainability. First and foremost, the healthiest animals will also be the most productive and economically efficient animals. Thus, prioritizing animal health and well-being in AI-assisted proactive livestock production systems benefits both farmers and the companies providing various sensors and decision support tools.

Proactive livestock production requires a system-level thought process. A farmer cannot manage what he or she cannot measure. In the example of a cough incidence

sensor, the goal is to identify a respiratory illness many days before a manager would have noticed the health issue. Then, the manager would intervene in a timelier manner to ultimately reduce mortality and improve animal performance. While in some cases increased coughing results in increased mortality, it is not a perfect indicator of the outcome variable mortality, which is most important to the farmer. Mortality in pigs is highly variable and influenced by hundreds of factors, such as temperature, age, and genetics, and cough incidence is just one piece of the puzzle (Gebhardt et al., 2020a; Gebhardt et al., 2020b). To manage and quantify direct effects of a factor on mortality rate, daily measurement of mortality is necessary. To accurately forecast changes in mortality, a diverse set of data streams comprising sensors, cameras, and production data is necessary. This holds true for many other key performance indicators outside mortality. However, as the number of data streams increases, the decision to modify management protocols becomes more complex. Incorporation of these data streams into a machine learning forecasting model solves the issue of complexity and removes human error in the decision-making process.

Comprehensive, well-designed training, validation, and testing frameworks are necessary for machine learning models that have sufficient accuracy to enable proactive management and information rich genetic prediction (Figure 1.1). Machine learning models that are successfully deployed can forecast deviations from normal for a given outcome variable, such as mortality, days in advance. With adequate warning, managers can devise intervention plans or consult veterinarians and anticipate increases in mortality or changes in other key outcomes. Continued research in sensors, cameras, and other automated high-throughput data streams are critical to provide a wider range of indicators

for production outcomes. Further, a systems perspective allows key biological variables, such as basal metabolic rate from cameras or sensors, to be predicted in genetic evaluations, rather than indicators, such as body weights. In addition, time investment in the collection of daily production data remains important to maximize the impact of machine learning models. In the face of labor shortages, prioritizing the collection of the most impactful data that characterize animal health, system-wide efficiency, and direct economic impacts is critical to the application of these technologies in the coming years.

CONCLUSIONS

As innovation continues in livestock production, there will be an abundance of data and tools available to farmers and researchers to help manage livestock. Machine learning models and artificial intelligence systems enable a shift from reactive to proactive livestock production. Additional investment of time and financial resources to collect high quality data using sensors and cameras increases the impact of machine learning models on proactive livestock production. Adoption of artificial intelligence on farms and ranches will have a measurable positive impact on animal health and well-being. This approach also allows the measurement of more influential phenotypes to include in genetic evaluations. Ultimately, artificial intelligence enabled proactive management will improve the sustainability of livestock production in the face of a growing population and demand for high quality animal protein across the globe.

FIGURES

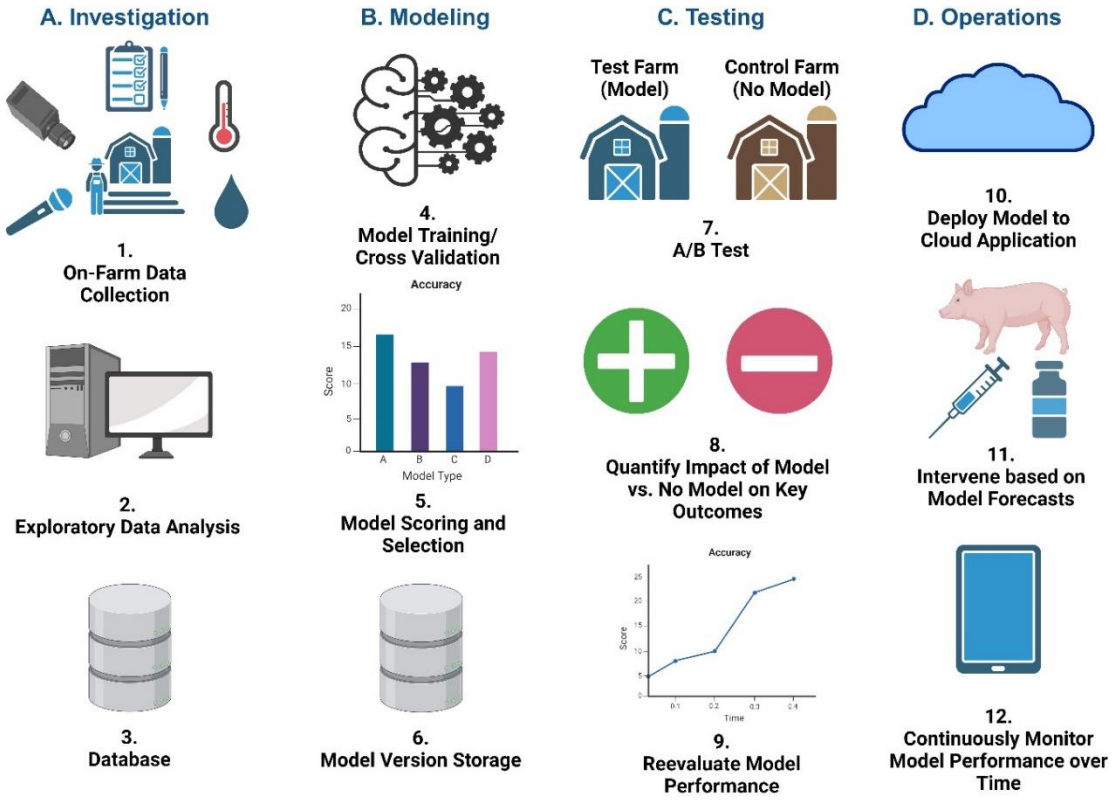


Figure 1.1. Comprehensive illustration of a proposed machine learning framework consisting of four key phases for proactive livestock management. **(A) Investigation Phase:** evaluation of sensors and production data for associations with key outcome variables. **(B) Modeling Phase:** curation of training dataset and evaluation of machine learning models for predictive performance using cross-validation. **(C) Testing Phase:** evaluation of best performing machine learning model in the production setting for impact on key outcome variables and management staff behavior. **(D) Operations Phase:** deployment of machine learning model across a production system and continuous monitoring of forecasting performance. Only models that have made measurable positive impacts in the testing phase should be deployed into production. Created with BioRender.com.

CHAPTER 2. ANALYSIS OF POLYGENIC SELECTION IN PUREBRED AND CROSSBRED PIG GENOMES USING GENERATION PROXY SELECTION MAPPING

Caleb J. Grohmann[†], Caleb M. Shull[ψ], Tamar E. Crum[¥], Clint Schwab[¥], Timothy J. Safranski[‡], and Jared E. Decker[†‡§]*

[†]Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA

[‡]Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA

[§]Genetics Area Program, University of Missouri, Columbia, MO 65211, USA

[ψ]The Maschhoff's, LLC, Carlyle, IL 62231, USA

[¥]AcuFast, Ltd., Saskatoon, SK S7K 2K6, Canada

*Corresponding author

ABSTRACT

Background

Artificial selection on quantitative traits using breeding values and selection indices in commercial livestock breeding populations causes changes in allele frequency over time, termed polygenic selection, at causal loci and the surrounding genomic regions. Researchers and managers of pig breeding programs are motivated to understand the genetic basis of phenotypic diversity across genetic lines, breeds, and populations using selection mapping analyses. Here, we applied Generation Proxy Selection Mapping (**GPSM**), a genome-wide association analysis of SNP genotype (38,294 to 46,458 markers) of birth date, in four pig populations (15,457, 15,772, 16,595 and 8,447 pigs per population) to identify loci responding to artificial selection over a span of five to ten years. Gene-drop simulation analyses were conducted to provide context to GPSM results. Selection signatures within and across each population of pigs were compared in the context of swine breeding objectives.

Results

Forty-nine to 854 loci were identified by GPSM as under selection (Q-values less than 0.10) across 15 subsets of pigs based on combinations of populations. The number of significant associations increased as populations of pigs were pooled. In addition, several significant associations were identified in more than one population. These results indicate concurrent selection objectives, similar genetic architectures, and shared causal variants responding to selection across populations. Negligible error rates (less than or equal to 0.02%) of false-positive associations were identified when testing GPSM on

gene-drop simulated genotypes, suggesting that GPSM distinguishes selection from random genetic drift in actual pig populations.

Conclusions

This work confirms the efficacy and negligible error rates of the GPSM method in detecting selected loci in commercial pig populations. Our results suggest shared selection objectives and genetic architectures across swine populations. Identified polygenic selection highlights loci important to swine production.

BACKGROUND

Broadly, population genetic methods are used to identify three types of directional selection in genomic data. First, hard selective sweeps are the signature of rapid selection in which one haplotype is selected to fixation within a population. Under this rapid selection, variation surrounding the selected mutation is drug or hitchhikes with the selected mutation resulting in large tracks of reduced nucleotide diversity and haplotype homozygosity. Soft selective sweeps are similar to hard sweeps in that diversity is reduced around the selected locus, but the selected DNA variants are on more than one haplotype, via selection on standing variation, recurrent mutation or migration (Hermisson and Pennings, 2005; Pennings and Hermisson, 2006). Finally, polygenic selection is a large change in a phenotype that results in small changes in allele frequency at hundreds or thousands of loci (Barghi et al., 2020).

Artificial selection in pigs, over the past 300 years, has led to the formation of pig breeds with well-defined breed characteristics and considerable across breed variation in phenotypes related to economically relevant traits (Wilkinson et al., 2013). Pig breeders placing selection pressure on certain qualitative phenotypes such as coat color and ear morphology and quantitative phenotypes such as feed efficiency, average daily gain, and backfat depth has left signatures of selective sweeps across the genomes of pig populations (Moon et al., 2015). In general, selective sweeps which are large, rapid changes in allele frequency which drag neighboring variation, leaving pronounced signatures of selection, are associated with phenotypes that underly the divergence of pig breeds, and have been identified in pig genomes by several studies (Wilkinson et al., 2013; Yang et al., 2014; Moon et al., 2015). However, pig breeders are more concerned

with selection for increased rates of genetic gain in quantitative traits (Ibáñez-Escriche et al., 2014), which are influenced by hundreds or thousands of genes. Further, the selection index has been the preferred method to improve the aggregate genetic merit of pigs by combining data from multiple quantitative traits (Hazel, 1943; Hazel et al., 1994), further increasing the number of genes under selection. Artificial selection using selection indices in pig breeding programs has been proven to cause significant changes to the mean phenotype of any one trait considered within the breeding objective (Hazel and Lush, 1942; Ellis et al., 1988; Stas, 2017). However, artificial selection pressure, especially over relatively short time scales, causes only subtle changes to allele frequencies at quantitative trait loci (**QTL**) across the genome (Kessner and Novembre, 2015; Rowan et al., 2021). In addition, loci that affect traits that are not explicitly included in the selection index, such as innate immunity, have been implicated to undergo frequency changes as a result of selection pressure applied in livestock breeding programs (Decker et al., 2012; Rowan et al., 2021).

There is much interest within livestock genomics in deciphering the genetic basis of phenotypic diversity in species raised for meat production (Decker et al., 2012; Rowan et al., 2021). Understanding selection in livestock populations is of paramount importance when evaluating the genomic basis of phenotypic variation within a genetic line, breed, or entire livestock population over time. The identification of selection detects loci that have been subjected to consistent increases or decreases in allele frequency significantly larger than due to random genetic drift (Kreitman, 2000; Gouveia et al., 2014; Gurgul et al., 2018). Unlike hard or soft sweeps, polygenic selection does not leave distinctive signatures on the genome (Rowan et al., 2021). With current

technologies such as single nucleotide polymorphism (**SNP**) arrays, temporally distributed genotypes, and increased computing resources, statistical analysis of polygenic selection is now feasible. Identification of regions of the genome that have been altered due to artificial selection pressure is highly beneficial in ascertaining QTLs under selection (Rowan et al., 2021). When results of selection mapping analyses are combined with results from phenotype-based genome-wide association studies (**GWAS**), QTLs associated with phenotypic variation of traits within breeding objectives can be supported from multiple lines of evidence (Qanbari and Simianer, 2014). Moreover, there are opportunities within selection mapping analyses to evaluate results within or across genetic lines or breeds, which can highlight differences in selection objectives across livestock breeding programs. Selection mapping analyses are not limited to increasing knowledge with respect to selection and evolution of species. Further, using results from selection analyses, SNP assays used for genomic prediction of breeding values in livestock populations can be refined in order to reduce extraneous statistical noise and increase prediction accuracy. This prioritization of SNPs can be accomplished by excluding SNPs that have not undergone significant changes due to directional selection or have not contributed to genetic change in traits in the breeding objective.

Generation Proxy Selection Mapping (**GPSM**) has been used as an analytical method for detection of polygenic selection loci in populations (Decker et al., 2012; Walsh et al., 2018; Rowan et al., 2021). In this approach, animal birth date (or other generation proxy) is fit as the dependent variable, and SNPs that are strongly associated with birth date are identified. If a SNP is under directional selection pressure, changes in allele frequency will generally be consistent over time, and an animal's genotype will be

strongly associated with birth date (Decker et al., 2012; Rowan et al., 2021). In addition, a major advantage in the applicability of GPSM methodology to livestock species over other methods, such as site frequency spectrum and linkage disequilibrium-based methods (Weigand and Leese, 2018), is the ability to adjust for demography and confounding due to non-random ascertainment of genotype samples, population structure, inbreeding, or kinship with the use of a genomic relationship matrix (**GRM**) (Decker et al., 2012; Rowan et al., 2021). Generation proxy selection mapping has been proven effective and accurate in identifying loci with changes in allele frequency due to polygenic selection (as opposed to loci-specific allele frequency changes due to random genetic drift) in beef cattle populations that have been exposed to artificial selection for approximately 50 years (Rowan et al., 2021). However, there are stark differences between beef breeding programs and swine breeding programs. For example, generation intervals in pigs are much shorter than in cattle (2 to 2.5 versus 4 to 5 years, respectively) (Jonas and Koning, 2015). Thus, for traits with similar accuracy and assuming similar selection intensity, comparable amounts of genetic gain are expected in approximately half the time for pig populations versus beef cattle populations. Moreover, due to increasing adoption of specialized sire and dam lines, the classical “breeding pyramid”, and vertical integration in the swine industry, breeding objectives within a population of pigs tend to be more focused than breeding objectives within beef breeds, where each breeder and farm have their own breeding objectives that may be poorly defined. The described differences between cattle and swine breeding programs contribute to variation in the effect of artificial selection on allele frequencies over time. The objectives of the current study were to 1) use GPSM to identify loci under artificial selection in three

purebred populations and one crossbred population of pigs and 2) compare and contrast the effect of artificial selection patterns among genotypes of each population in the context of a swine breeding company.

METHODS

Population Background

Four populations of pigs were used in the present study (data owned by The Maschhoff's, LLC, Carlyle, Illinois, USA). Within each population, a selection index was utilized to identify boars and gilts with superior genetic merit to return to the breeding population at the nucleus level. Breeding population-specific selection indices for all populations included expected progeny differences (**EPDs**) for growth and carcass traits such as increased feed efficiency and average daily gain, decreased backfat depth, and increased *Longissimus* muscle area. In addition, selection indices for two of the four breeding populations (Landrace and Yorkshire) also emphasized maternal reproductive traits and included EPDs for increased number and weight of piglets born and weaned.

Pedigree and Genotype Data

A pedigree consisting of individual, sire, and dam identification, birth date, and genetic line for 1,247,982 pigs was provided by The Maschhoff's. Summary information regarding the number of sires and dams, founder pigs, and generations within each population is presented in Table 2.1. On a subset of 16,802, 19,342, 18,368, and 8,532 pigs from the Duroc, Landrace, Yorkshire, and Crossbred populations, respectively, SNP assays were collected using a GGP Porcine 50K (Neogen, Corp., Lansing, Michigan, USA) genotyping array. Genomic coordinates for each SNP were from the Sscrofa 11.1 reference genome (Warr et al., 2020). Sample collection and subsequent genotyping was

conducted on all viable male selection candidates prior to removal from performance testing trials. In addition, all female animals selected to return to the nucleus breeding herd were genotyped. Summary information regarding the number of sires and dams, founder pigs, birth date ranges, and generations for genotyped pigs within each population is provided in Table 2.2.

Preparation of Genotype Data and Overview of Analyses

The dependent variable for all analyses was birth date (**AGE**) calculated as the difference, in months, between each pig's birth month and January 2006. Pigs from the entire dataset of genotyped pigs were separated into 15 subsets based on population or combination of populations. Analyses were conducted using only SNPs located on the autosomal chromosomes for *Sus scrofa,* which were chromosomes 1 through 18. Genotype quality control was performed in PLINK v1.9 (Purcell et al., 2007) for each subset. Any SNP with a genotype call rate less than 0.90 or a minor allele frequency less than 0.01 was filtered from the data. In addition, individual pigs that had a genotype call rate less than 0.90 were filtered from the dataset.

Percent Duroc, Landrace, and Yorkshire ancestry was predicted for each pig using fastSTRUCTURE (Raj et al., 2014), with the *K* parameter set to 3. Purebred pigs that were predicted to be less than 95% of their assigned genetic line (Duroc, Landrace, or Yorkshire) were removed from all subsequent analyses, as these may represent sample swaps. While predicted breed proportions were estimated for the Crossbred pigs, none were removed from the genotyped sample, as deviations from expected breed proportions cannot be distinguished from deviations due to Mendelian sampling or noise of ancestry prediction. Genomic relationship matrices (**GRMs**) were estimated for each subset using

the software GCTA v1.93.2 (Yang et al., 2011) and the method described by Yang et al. (Yang et al., 2010), and these GRMs were utilized in all subsequent analyses. To visualize the genomic relatedness between lines, the 'pca' function of GCTA (Price et al., 2006) was also used to conduct a principal component analysis (**PCA**) on a GRM for all Duroc, Landrace, Yorkshire, and Crossbred pigs. A summary of the number of pigs and SNPs after quality control and all subsequent analyses performed for each subset is presented in Table 2.3. Descriptive statistics of AGE by genetic line were calculated using the 'dplyr' package (Wickham et al., 2023) of the statistical analysis software R (R Core Team, 2023). Figures were generated using the 'ggplot2' (Wickham, 2016) package of R.

Depending on data subset, certain combinations of the following three statistical analyses were performed on AGE: 1) univariate variance component estimation, 2) bivariate variance component estimation, and 3) univariate genome-wide association using a mixed linear model to estimate SNP associations.

Univariate Variance Component Estimation

To estimate the proportion of variance in AGE explained by genome-wide SNPs (**PVE**) for each subset (Table 2.3), the following model was fit using GCTA:

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{Zu} + \boldsymbol{e}$$

$$\boldsymbol{u} \sim N(0, \boldsymbol{G}\sigma_g^2)$$

$$\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$$

where **y** is a vector of observations for AGE, μ is the overall mean for AGE, **u** is a vector of random polygenic effects, **Z** is an incidence matrix relating AGE in **y** to random polygenic effects in **g**, and **e** is a vector of random residuals, **G** is the genomic

relationship matrix, and **I** is an identity matrix. Additive genetic (σ_g^2) and residual (σ_e^2) variance components were derived using average information restricted maximum likelihood. The PVE was then estimated as follows:

$$PVE = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

Bivariate Variance Component Estimation

Genetic correlations ($\mathbf{r_G}$) between each population (Table 2.3) for AGE were estimated using bivariate mixed linear models, fitted in GCTA, of the following form:

$$\begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{bmatrix} = [\mathbf{1}] \begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z_1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{g_1} \\ \boldsymbol{g_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix}$$

where $\boldsymbol{y_1}$and $\boldsymbol{y_2}$ are vectors of observations for AGE for the two populations, μ_1 and μ_2 are the overall means for AGE for each population, respectively, $\boldsymbol{g_1}$and $\boldsymbol{g_2}$ are vectors of random polygenic effects for each pig in the two populations, $\boldsymbol{e_1}$and $\boldsymbol{e_2}$ are residuals for AGE of the two populations, and $\boldsymbol{Z_1}$and $\boldsymbol{Z_2}$ are incidence matrices for the random polygenic effects in $\boldsymbol{g_1}$ and $\boldsymbol{g_2}$, respectively. Additive genetic variance of $\boldsymbol{g_1}$ and $\boldsymbol{g_2}$ ($\sigma_{g1}^2$and σ_{g2}^2, respectively), additive genetic covariance between $\boldsymbol{g_1}$ and $\boldsymbol{g_2}$ ($\sigma_{g1,g2}$), and residual variance of $\boldsymbol{e_1}$ and $\boldsymbol{e_2}$ ($\sigma_{e1}^2$and σ_{e2}^2, respectively) were estimated using average information restricted maximum likelihood with the variance-covariance matrix (**V**) defined as:

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{Z_1 G Z_1' \sigma_{g_1}^2 + I \sigma_{e_1}^2} & \boldsymbol{Z_1 G Z_2' \sigma_{g_1 g_2}^2} \\ \boldsymbol{Z_2 G Z_1' \sigma_{g_1 g_2}^2} & \boldsymbol{Z_2 G Z_2' \sigma_{g_2}^2 + I \sigma_{e_2}^2} \end{bmatrix}$$

where **G** and **I** were the genomic relationship and identity matrix, respectively. Genetic correlations were then estimated by GCTA using the following formula:

$$\boldsymbol{r_G = \frac{\sigma_{g1,g2}}{\sqrt{\sigma^2_{g1} \cdot \sigma^2_{g2}}}}$$

Generation Proxy Selection Mapping (GPSM)

Generation Proxy Selection Mapping analyses were conducted to detect SNPs with allele frequency changes over time within each subset (Table 2.3). To accomplish this, single-SNP univariate mixed linear models were fit in GCTA as part of GWAS of AGE, with the models defined as follows:

$$\boldsymbol{y = Xb + Zg + e}$$

$$\boldsymbol{g} \sim N(0, \boldsymbol{G}\sigma^2_g)$$

$$\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma^2_e)$$

where **y** is the pig's generation proxy (AGE), and **X** is an incidence matrix that relates SNPs to AGE for each pig and **b** is the estimated SNP effect. Confounding due to population structure, relatedness, and inbreeding are controlled by the random polygenic term **g**, and **Z** is an incidence matrix for the effect **g**. In addition, **G** is the genomic relationship matrix, and **I** is an identity matrix. Additive genetic (σ^2_g) and residual (σ^2_e) variance components were estimated using average information restricted maximum likelihood. However, these variance components were not of interest as a part of the GPSM analyses as they were estimated previously as a part of the univariate variance component estimation analysis. *P*-values of estimated SNP effects were converted to false discovery rate (**FDR**) corrected *Q*-values using the 'qvalue' package (Storey et al., 2017) of R, and a significance threshold of $Q < 0.10$ was used for all analyses.

Variance Component and GPSM Analyses using Simulated Data

Variance component and GPSM analyses of purebred pigs (subsets 1, 2, 3, 5, 6, 8, and 11; Table 2.3) were conducted using gene-drop simulated genotype data produced using the package 'AlphaSimR' (Gaynor et al., 2020; Gaynor et al., 2021) with the pedigrees of the analyzed pigs. The objective of these gene-drop analyses was to ensure GPSM results performed on real data were due to artificial selection as opposed to random genetic drift. For the univariate variance component estimation and GPSM analyses for each of the subsets 1 through 3, 5, 6, 8, and 11 (Table 2.3), 5,000 founder pig haplotypes were simulated using AlphaSimR's MaCS (Chen et al., 2008) wrapper, with the demography parameter set to "GENERIC". Each of the simulated haplotypes contained 90,000 segregating sites located evenly across 18 chromosomes. Then, using the pedigreeCross function (Gaynor et al., 2020), founder pigs in the pedigree of each subset were assigned genotypes at random from the simulated population of 5,000 pigs. Simulated founder pig haplotypes were then dropped through each pedigree to simulate the exact matings that have occurred in The Maschhoff's breeding program (allele inherited by progeny was randomly assigned according to recombination and segregation). Lastly, pigs with genotypes used in the real analyses were pulled from each subset along with a "SNP chip" of randomly selected loci equivalent to the number of SNPs used in the real analyses (Table 2.3). Univariate variance component estimation and GPSM analyses were conducted using the same statistical models and software, the simulated genotypes, and the AGE values from the real analyses. For the Duroc, Landrace, and Yorkshire populations, the above process was replicated five times to

ensure results from analysis of simulated data were not affected by randomness within the simulation process.

In bivariate variance component analyses on simulated data (subsets 5, 6, and 8; Table 2.3), founder pig haplotypes were simulated two different ways. First, founder pig haplotypes were simulated as one group that consisted of 15,000 founder pigs (Method 1). The objective of this method was to simulate a scenario where each combination of populations had recently diverged; thus, the founder animals for each population have the same genotypes. For the second method, founder pig haplotypes were simulated separately for each population, the random number generator in R was changed between each simulation, and then the two founder pig haplotypes were combined (Method 2). Using Method 2, the simulated genotypes were vastly different between founder pigs in each population combination, which represented pairs of populations that were completely unrelated. These two strategies represent the extremes of coalescent times between breeds, rather than assuming a specific number of generations since the divergence of the breeds. Samples of pigs with simulated genotypes were created in the same manner as described above for the univariate analyses. Bivariate variance component analyses were then conducted using both samples of simulated genotypes from each method and each pairwise comparison of subsets 5, 6, and 8. Results from all analyses using simulated data and analyses using real data were then compared in a one-to-one fashion.

Investigation of GPSM Associations

The number of shared significant GPSM associations between and across each purebred population and the crossbred population (subsets 1 to 4; Table 2.3) were

visualized using the R package 'UpSetR' (Conway et al., 2017). The 'GALLO' package of R (Fonseca et al., 2020) was used to identify positional candidate genes [file Sus_scrofa.Sscrofa11.1.105.gtf.gz downloaded from the 'Pig' section of Ensembl (Cunningham et al., 2022)] and quantitative trait loci [file Animal_QTLdb_release76_pigSS11.gff.gz downloaded from the 'PigQTLdb' section of AnimalQTLdb (Hu et al., 2022)] within 100 kb upstream and downstream of each significant SNP identified by GPSM in the Duroc, Landrace, Yorkshire, and Crossbred populations (subsets 1 to 4; Table 2.3) In addition, the 'gwascat' package of R (Carey, 2023) was used to download the most recent version of the NHGRI-EBI GWAS Catalog (Sollis et al., 2023). Traits from the NHGRI-EBI GWAS Catalog that were located within genes annotated by 'GALLO' were identified and discussed.

RESULTS

Descriptive Statistics and Principal Component Analysis

Descriptive statistics of AGE for each subset are presented in Table 2.4. In addition, raw distributions of AGE are shown in Figure 2.1 for all pigs (subset 15; Table 2.3) and Figure 2.2 for subsets 1 through 4 (Table 2.3). The histograms of AGE depict the frequency of genotype sampling across all populations and within each population for the duration of The Maschhoff's breeding program (Figures 2.1 and 2.2, respectively). In general, descriptive statistics for AGE were similar across each subset (Table 2.4). However, the range and standard deviation of AGE for the Crossbred pigs was less than that of the other subsets, as genotyping for these pigs did not begin until March of 2015 (Table 2.2). Thus, the number of genotyped Crossbred pigs was approximately half of the number of Duroc, Landrace, and Yorkshire pigs. Furthermore, histograms of AGE for

each subset were left-skewed, indicating that the number of pigs genotyped per year in each subset generally increased from the start of The Maschhoff's SNP collection platform in 2010 until 2020.

Results from the PCA of the GRM containing all pigs from each population (subset 15; Table 2.3) are presented in Figure 2.3. By plotting principal component 1 versus principal component 2 for the genomic relatedness of these four populations, four defined clusters were visualized, as expected. In addition, the cluster for the Crossbred population was located halfway between the Duroc population cluster and the Landrace and Yorkshire population clusters along principal component 1 and halfway between the Landrace and Yorkshire population clusters along principal component 2 (Figure 2.3). McVean postulated that the location of an admixed population of individuals on a PCA plot relative to its source populations directly relates to the admixture proportion of these individuals amongst the source populations (McVean, 2009). Thus, the location of the Crossbred cluster in Figure 2.3 confirms approximately a 50/25/25 admixture amongst the Duroc, Landrace, and Yorkshire populations, respectively. This result was to be expected given the design of The Maschhoff's mating program for their commercial test herd, which mates Duroc sires to Landrace × Yorkshire dams.

Univariate and Bivariate Variance Component Estimation

The proportion of variation explained by genome-wide SNPs of the dependent variable AGE for all 15 subsets is presented in Table 2.5. These values ranged from 0.81 to 0.94 (Table 2.5) and were significantly greater than zero ($P < 0.001$) using the likelihood ratio test. Previous GPSM simulations have shown that GPSM PVE is indicative of population demographic history (Rowan et al., 2021). Given that descriptive

statistics and distributions of AGE were generally similar across subsets (Table 2.4; Figures 2.1 and 2.2), these PVE results suggest that the four populations have similar demographics, such as inbreeding, effective population sizes, and pedigree structure. Furthermore, the results of univariate variance component estimation in subsets containing purebred populations (subsets 1, 2, 3, 5, 6, 8, and 11; Table 2.3) using simulated genotypes are presented in Table 2.6. Estimated PVEs for subsets with simulated genotypes were generally like those with real data, ranging from 0.83 to 0.93 (Table 2.6), and were significantly different from zero ($P < 0.001$) using the likelihood ratio test. Between the univariate variance component estimation analyses using real and simulated genotype data, the pedigree structure, AGE values, and number of SNPs were the same for each subset; however, the genomic relationship between each pairwise combination of pigs was different. For the Landrace and Yorkshire populations, the real PVE was only higher than the simulated PVE by 0.02 and 0.01, respectively. However, for the Duroc population, the real PVE was 0.11 higher than the simulated PVE. In addition, PVEs from univariate variance component analyses were similar across replicates of simulated genotype data. For example, PVEs ranged from 0.82 to 0.84 (mean of 0.83 ± 0.004), 0.85 to 0.87 (mean of 0.86 ± 0.003), and 0.84 to 0.85 (mean of 0.842 ± 0.0020) for the Duroc, Landrace, and Yorkshire populations, respectively, across five replications of simulated genotype data per population (Supplementary Table 2.1). Thus, the variance components and PVE are not impacted by stochastic generation of the simulated genotypes.

Genetic correlations between AGE for all pairwise combinations of the Duroc, Landrace, Yorkshire, and Crossbred populations (subsets 1 through 4; Table 2.3) are

presented in Table 2.7. In general, genetic correlations between purebred populations were stronger than genetic correlations between each purebred population and the Crossbred population (Table 2.7). Each genetic correlation was significantly different from zero ($P < 0.001$) using the likelihood ratio test. Within the purebred subsets (5, 6, and 8; Table 2.3), the genetic correlation between Landrace and Yorkshire pigs was higher than the genetic correlation between Duroc and Landrace or Duroc and Yorkshire pigs (Table 2.7). This indicates that the demography and selection (associations with AGE) are more similar between Landrace and Yorkshire pigs than between either of the two maternal breeds and the Duroc population. This result was expected, as Landrace and Yorkshire pigs are both selected for maternal traits while Duroc pigs are selected for increased efficiency in terminal traits. Amongst each pairwise combination between the Crossbred population and each purebred population (subsets 7, 9, and 10; Table 2.3), genetic correlations for AGE were highest between the Duroc and Crossbred population and were similar between Crossbred and Landrace or Yorkshire pigs (Table 2.7). Given that Duroc pigs contribute more genetic material to the Crossbred pigs than the Landrace and Yorkshire pigs, this result was expected.

Table 2.8 presents genetic correlations from the bivariate variance component estimation analyses using simulated data. Genetic correlations between each population (subsets 5, 6, and 8; Table 2.3), using both methods, were not significantly different from zero ($P > 0.05$) using the likelihood ratio test (Table 2.8). This result suggests that in the absence of artificial selection pressure on economically relevant traits in each population, transmission of genotypes between generations is independent across breeds, hence the genetic correlation is expectedly zero. Moreover, miniscule genetic correlations were

observed across both methods used to simulate founder populations; therefore, the length of time from population divergence likely has no effect on genetic correlations in the presence of genetic drift. Thus, the results from this bivariate variance component analysis using simulated founder genotypes strengthen the validity of the assumptions presented above based on real genotype data in genetic lines exposed to artificial selection pressure.

Detecting Polygenic Selection with Generation Proxy Selection Mapping (GPSM)

The number of significant SNPs ($Q < 0.10$) associated with AGE for each subset is presented in Table 2.9. Although the distribution of AGE for each subset was left-skewed and non-normal (Table 2.4; Figures 2.1 and 2.2), the GPSM *P*-values for independent SNP genotype association tests with AGE were well calibrated (Figure 2.4). For example, *P*-values for null SNPs, which were deemed non-significant by GPSM, closely followed the expected uniform distribution, while SNPs that were significantly associated with AGE deviated from this expectation (Figure 2.4). This result suggests that departures from normality in the dependent variable in a GPSM analysis does not produce spurious associations between AGE and genotype. Generation proxy selection mapping identified 49 to 854 significant SNPs (Table 2.9) depending on subset. The number of significant associations generally increased as the number of samples in the subset increased, as expected, due to increased power of the GWAS.

There were 100, 147, 138, and 49 significant SNPs identified by GPSM representing 0.26, 0.33, 0.31, and 0.11% of the total number of autosomal SNPs for the Duroc, Landrace, Yorkshire, and Crossbred populations, respectively (subsets 1 through 4; Table 2.9). However, when all purebred pigs were combined into a single subset

(subset 11; Table 2.3), GPSM identified 702 significant associations (1.51% of autosomal loci; Table 2.9). Moreover, the addition of the Crossbred pigs to subset 11, which created subset 15 (Table 2.3), allowed GPSM to identify 854 significant associations (1.84% of autosomal loci; Table 2.9). As mentioned above, the efficacy of GPSM analyses depends on the power of the genome-wide association analyses. Thus, as more samples of SNP genotype information on a particular population of pigs are accumulated, more SNP genotypes that are associated with AGE can be detected using the GPSM method.

Manhattan plots of $-log_{10}(Q)$ values for the associations between SNP genotypes and AGE in the Duroc, Landrace, Yorkshire, and Crossbred populations (subsets 1 through 4; Table 2.3) are presented in Figure 2.5. For each population, a plot is presented with a full (Figure 2.5A-2.5D) and a truncated Y-axis, which ranged in $-log_{10}(Q)$ values from 0 to 10 (Figure 2.5E-2.5H). Within each subset, several significant associations between SNP genotype and AGE were identified on each chromosome by GPSM (Figure 2.5). When viewing the Manhattan plots for each genetic line with truncated Y-axes, genome-wide nature of the significant associations becomes more pronounced (Figure 2.5E-2.5H).

Distributions of SNP effects are plotted in Figure 2.6. The SNP effects in each population that were significantly different than zero were converted to absolute values to interpret differences in magnitude of this parameter across populations. Duroc pigs had the highest mean absolute value of age SNP effects for significant SNPs (2.70 months) and mean absolute values of SNP effects in significant SNPs were similar among the Landrace (1.66 months), Yorkshire (1.55 months), and Crossbred (1.80 months) populations. The range in absolute values of SNP effects in GPSM significant SNPs was

considerable, depending on population. For example, in the Duroc and Crossbred populations, these ranges were 1.00 to 13.32 months and 0.70 to 14.39 months, respectively. However, for the Landrace and Yorkshire populations, these ranges were smaller (0.71 to 6.64 and 0.71 to 6.00 months, respectively) but were similar between the two populations. In addition, the mean change in allele frequency per year in significant SNPs for each population was 0.018 per year for Duroc (range 0.00001 to 0.109), 0.019 per year for Landrace (range 0.0001 to 0.082), 0.019 per year for Yorkshire (range 0.0006 to 0.101), and 0.024 per year for Crossbred (range 0.0007 to 0.086).

Results from GPSM analyses using randomly simulated founder genotypes are presented in Table 2.10. Out of the 11 GPSM runs on the simulated data, GPSM falsely identified significant associations with AGE in seven analyses (Table 2.10). However, in these analyses, a very small number of spurious associations were detected (Table 2.10), corresponding to a range of error rates between 0 to 0.0152% (Table 2.10), which are negligible. Moreover, the false positive rate for GPSM associations was stable across five replicates of simulated genotype data for the Duroc, Landrace, and Yorkshire populations (mean of 0.0031 ± 0.00192%, 0.0022 ± 0.00121%, and 0.0004 ± 0.00044%, respectively; Supplementary Table 2.2).

Analyses that used Method 2 to simulate founder genotypes, which simulated completely different founder genotypes for each population, had a significantly higher number of spurious associations than Method 1, which simulated a single founder population for all three purebred populations (paired t-test, t = -4.2748, df = 3, *P*-value = 0.0235). In commercial pig populations, however, divergence likely happened in a scenario that resembles a blending of Methods 1 and 2; thus, the higher error rate in

Method 2 could be inflated compared to reality in the swine industry. Nevertheless, these results suggest that GPSM is robust to allele frequency changes due to genetic drift over time.

The number of shared significant GPSM SNPs across subsets 1 through 4 is presented in Figure 2.7. Forty-two, 22, and 4 SNPs significantly associated with AGE were shared across at least two, three, or four populations, respectively (Figure 2.7). Twenty-five GPSM associations were shared between the Landrace and Yorkshire populations, which was considerably larger than the number of shared GPSM associations identified between all other pairwise combinations of population (Figure 2.7). In addition, 13 GPSM associations were shared across all three purebred populations. However, only 2 to 4 GPSM SNPs were unique to subsets of three populations where the Crossbred pigs were included (subsets 12 through 14; Figure 2.7). Top SNP associations with AGE in the Duroc, Landrace, Yorkshire, and Crossbred populations are shown in Table 2.11. In general, most of the significant SNPs with the 10 largest absolute values for SNP effects were significant in at least one other subset (Table 2.11). In the Crossbred population, only 2 of the top 10 large effect SNPs were unique to Crossbred pigs (3, 4, and 1 out of 10 were significant across at least 2, 3 and 4 subsets, respectively; Table 2.11). In addition, certain SNPs exhibited large SNP effects across multiple subsets. For example, GPSM estimated a SNP effect for SNP 39502 of 6.19, -6.64, and 4.03 months (Table 2.11) in the Duroc, Landrace, and Yorkshire populations, respectively, which were 11.9, 18.4, and 10.9 SD above, below, and above the mean SNP effect within each population, respectively.

Supplementary Table 2.3 contains all positional candidate genes and quantitative trait loci in identified in pigs [AnimalQTLdb (Hu et al., 2022)] and humans [(NHGRI-EBI GWAS Catalog (Sollis et al., 2023)] located within 100 kb upstream or downstream of GPSM significant SNPs in the Duroc, Landrace, Yorkshire, and Crossbred populations. Eight positional candidate genes were identified in all four populations. Specifically, *STX11* and *UTRN* were identified on chromosome 1, *AP3B2*, *FSD2*, *HOMER2*, and *WHAMM* on chromosome 7, and *TMEM132D* and *U6* on chromosome 14. Moreover, fourteen positional candidate genes were identified in the Duroc, Landrace, and Yorkshire populations. More specifically, *PRKN* on chromosome 1, *GALNT17* on chromosome 3, *CRSP2* and *ZDHHC17* on chromosome 5, *U6* on chromosome 6, *DDIT4L* and *EMCN* on chromosome 8, *ASNS*, *DGKB*, *GLCCI1*, *MIOS*, *RELT*, *UMAD1* on chromosome 9, and *PLD5* on chromosome 10 were located within 100 kb upstream or downstream of significant GPSM SNPs in each of the three purebred populations.

DISCUSSION

Polygenic selection on quantitative traits, induces small changes in allele frequencies at numerous loci across the genome over time (Höllinger et al., 2019; Barghi et al., 2020; Rowan et al., 2021). Detection of this polygenic selection due to artificial selection over time for traits with complex architectures was the focus of this study. The increasing abundance of genomic information from SNP arrays (Decker, 2015) has allowed many researchers to study changes in genotypic and allelic frequency in commercial and indigenous global pig populations (Ai et al., 2013; Yang et al., 2014; Ma et al., 2015; Moon et al., 2015; Gurgul et al., 2018). The rapid increase in studies in this area has given rise to new analytical methods to detect large and small selection

signatures across in the genome of commercially reared livestock species, such as Generation Proxy Selection Mapping (GPSM) (Decker et al., 2012; Rowan et al., 2021). In the present study, GPSM was used to estimate variance components and SNP genotype associations with the dependent variable AGE, which was calculated as the difference in months from January 2006 in a large commercial population of pigs that was comprised of three distinct pure populations (Duroc, Landrace, and Yorkshire) and a crossbred population comprised of the three pure populations. We found that the genomic relationship matrix accounted for confounding due to pedigree and population structure consistently across seven gene drop simulations, as false positive rates ranged between 0% to 0.015% (Table 2.10 and Supplementary Table 2.2).

The proportion of variation in age explained by the GRM ranged from 0.81 to 0.94. Rowan et al. stated, based on of simulations of genotypes from random mating versus selection analyzed with GPSM, that PVE is a function of the number of generations of selection, the number of total crosses per generation, and the genotype sampling scheme (even or uneven across generations) (Rowan et al., 2021). Our results are consistent with these conclusions as similar results across univariate variance component analyses using real and simulated genotype data indicated pedigree structure and the distribution of AGE were the main determinants of PVE (Tables 2.5 and 2.6). The difference between the gene-drop simulation PVE and observed PVE was small for the Landrace and Yorkshire population but was 0.11 for the Duroc population. The main difference between simulations and observed data was the presence of selection, suggesting that the Duroc population was under stronger selection compared to the Landrace and Yorkshire populations. Selection indices for Duroc terminal populations

generally consist of only traits related to growth, carcass and feed consumption, while selection indices for maternal lines consist of the previously stated traits and additional traits related to maternal prolificacy. Selecting on more traits means slower change for individual traits and their causal variants, especially when these traits are lowly heritable and require large amounts of data for accurate genetic evaluations. Thus, overall genetic merit likely improved at a slower pace in the maternal populations compared to the Duroc population as a result of the added traits in the selection index. Further, Rowan et al. found smaller PVEs across three cattle populations [PVE = 0.52, 0.59, and 0.46 in Red Angus (n = 15,295), Simmental (n = 15,350), and Gelbvieh (n = 13,031) populations, respectively] of similar sample sizes to the purebred populations in the current study (Rowan et al., 2021). Differences between cattle and pigs in overall structure of the genetic selection programs related to the above factors likely contributed to the large difference between the PVEs reported by Rowan et al. and the present study (Rowan et al., 2021).

The estimation of genetic correlations between pairwise combinations of the Duroc, Landrace, Yorkshire, and Crossbred populations confirmed our assumptions of similarity (or dissimilarity) between populations in their demographic and selection histories. A genetic correlation between AGE of two populations near 1 suggests a high proportion of autosomal loci that are statistically associated with AGE undergoing similar changes in allelic frequency over time, with a genetic correlation between 0 and -1 suggesting the contrary (dissimilar or antagonistic changes in allele frequency in SNPs associated with AGE over time). The results of the simulation analysis, where randomly generated founder pig SNP genotypes are randomly dropped through the real pedigree of

each population (mimicking genetic drift), validate this assumption, as genetic correlations between populations were not significantly different from zero (regardless of the most recent common ancestor in simulations) using the likelihood ratio test ($P > 0.05$; Table 2.8). Selection objectives within The Maschhoffs are highly similar between the Landrace and Yorkshire populations and are the most dissimilar between the Duroc and each of the Landrace and Yorkshire populations. Estimated genetic correlations in the present study followed this premise, as the estimated genetic correlations between the two maternal breeds was higher than those estimated between the Duroc population and either the Landrace or Yorkshire populations (Table 2.7). However, across all four populations, the genetic correlations were significantly larger than zero, indicating that selected loci are similar across populations. This is supported by GPSM associations, as most strong associations were identified in multiple populations (Figure 2.7) and there was a general increase in associations when pooling populations (subsets 5 through 15; Table 2.9).

In our study, GPSM identified hundreds of SNPs significantly associated with AGE ($Q < 0.10$) in most populations (Table 2.9). There was a wide range in the number of pigs in each subset used in the GPSM analyses (Table 9). The GPSM method, as stated above and in other studies (Decker et al., 2012; Rowan et al., 2021), is a genome-wide association analysis, which are more powerful in the detection of SNP genotypes associated with a particular phenotype as the number of samples in the population increases, due to increased precision in estimating SNP effects at a particular marker (Decker, 2015). This inherent attribute of GWAS contributed to the vast differences in the number of significant associations between SNP genotypes and AGE across subsets, as

the number of significant SNPs showed a general increase with sample size (Table 2.9). However, this is only the case if the same loci are increasing in frequency across the different populations. The overwhelming majority of autosomal SNPs for each subgroup were not associated with AGE, according to GPSM results (98.2 to 99.9% of the autosomal loci; Table 2.9). However, GPSM detected several SNPs significantly associated with AGE on each chromosome (Figure 2.5). In addition, the nature of the genome-wide associations with AGE indicates that selection in these populations is likely polygenic (Figure 2.5E-2.5F).

The distribution of ages of the genotyped samples affects the power (false negative rate) of GPSM analyses, with more even sampling providing more power and uneven sampling decreasing power (Rowan et al., 2021). The ages of the genotyped samples in this swine data are more evenly sampled than many of the analyzed cattle datasets (Rowan et al., 2021). This may explain why we identified a relatively large number of selected loci with moderate SNP marker density. Our gene drop simulations, in agreement with the simulations of Rowan et al., show that uneven sampling across time has a negligible effect on false positive rates (Rowan et al., 2021).

A number of SNPs were detected by GPSM across at least two populations (Figure 2.7). Upon visual assessment of Manhattan plots of GPSM *Q*-values for each population, several regions along the autosomal genome that expressed similar patterns of GPSM significance across populations were identified (Figure 2.5). Of particular interest, from the chromosome 9 region associated with selection in all three purebred populations, the four candidate genes (*MIOS*, *RPA3*, *UMAD1*, and *GLCCI1)* are all differentially expressed in ovarian tissues (Li et al., 2017). Most notably, *MIOS,* which is

commonly referred to as the "missing oocyte gene", is well known for its role in meiosis regulation of oocyte development. In a study using *Drosophila,* a mutation in the *MIOS* gene caused erroneous oocyte development. Instead of stimulating progression through each stage of meiosis, the described mutation caused oocyte progression towards polyploid nurse cells as opposed to fully functional, mature haploid gametes (Iida and Lilly, 2004). While there are no known studies that have evaluated the impact of mutations in the "missing oocyte" gene in pigs, results from the present study suggest selection pressure in The Maschhoff's genetic program has had a significant effect on regions of the pig genome that influence fertility. As a litter-bearing species, pig breeders routinely place selection pressure on litter traits such as total born and number born alive, especially in Landrace and Yorkshire pig populations. In addition, not only does selecting young replacement animals influence allele frequencies at quantitative trait loci, but decisions on which animals to cull likely have similar effects. For example, gilts or sows in breeding populations that fail to express estrus cyclicity, conceive or farrow litters, or return to estrus within a reasonable period post-weaning are typically removed from the herd. Selection or culling of breeding animals due to reproductive performance and fertility issues, respectively, likely caused allele frequency changes at loci near these 4 genes on chromosome 9. In addition to the *MIOS* gene, two genes of particular interest, *HOMER2* and *WHAMM*, were identified near significant GPSM SNPs on chromosome 7 in all four populations. In humans, both *HOMER2* and *WHAMM* were associated with lung function. However, *HOMER2* was also associated with traits related to human body mass index. In addition, these two genes are located in regions of the pig genome associated with carcass traits in pigs such as backfat thickness, loin muscle depth and

area, carcass length, dressing percentage, and estimated carcass lean content. *HOMER2* and *WHAMM* were likely identified in each population due to the strong emphasis on carcass feed efficiency and lean meat production in selection indices for the Duroc, Landrace, and Yorkshire populations of the current swine breeding company. What is still unknown is whether *HOMER2* and *WHAMM* influence carcass traits through their effect on lung function (healthier pigs) or they have carcass-specific effects in swine. However, further quantitative trait association studies and bioinformatics analyses are required to test these alternatives. The locus containing *UTRN* and *STX11* had AnimalQTL annotations related to white blood cell counts and virus titers (immunity) as well as adiposity measures (production). In human GWAS, *UTRN* had an association with lung function (Kichaev et al., 2019) and *STX11* had an association with pre-treatment viral load in HIV-1 infection (Ekenberg et al., 2019). Interestingly, the combination of production effects and immunity effects may also affect this locus on chromosome 1, suggesting that loci affecting production and immunity might be common targets of selection across breeds.

The detection of significant associations across the autosomal genome in each of the Duroc, Landrace, Yorkshire, and Crossbred populations indicates artificial selection is influencing numerous genes in each of these populations of pigs. Further, to see an increase of power when pooling data across populations and shared signal across populations, there must be common causal variants (or at a minimum, casual genes) segregating in the populations, and the variants must be responding to similar selection objectives. Thus, concordant traits across selection indices for maternal and terminal pig

breeds are likely influenced by the same quantitative trait loci in the genome of each breed.

We confirmed GPSM is robust in separating changes in allele frequency due to genetic drift and artificial selection, through simulations. In each of the 11 gene-drop simulations, GPSM found very few spurious associations between SNP genotype and AGE (Table 2.10). Rowan et al. identified false positives as significant at a rate of one SNP per 100,000 tests, which was similar to the current study (Rowan et al., 2021).

Except for two outliers [SNP 30819 in the Crossbred population (14.39 months) and SNP 3420 in the Duroc population (13.32 months); Table 2.11], the absolute values of SNP effects ranged between 0 and 8.21 months. Mean absolute values for AGE SNP effects in significant associations were higher in the Duroc population (2.70 months) than the other two purebred populations (1.66 and 1.55 months for the Landrace and Yorkshire populations, respectively). This suggests that selection intensity is greater in the Duroc population, which induces larger changes in allele frequency over shorter periods of time than in the maternal breed populations. Selection in the Duroc population within The Maschhoff's has been focused on traits that increase the efficiency of terminal commercial progeny, such as increased growth and feed efficiency, decreased backfat depth, and increased carcass lean content. Growth and carcass traits, in general, have more accurate genetic predictions due to their moderate to large heritabilities, which increases selection response as opposed to selection on maternal traits such as number of piglets born alive and litter weaning weight (traits that are emphasized in The Maschhoff's Landrace and Yorkshire populations). Moreover, as stated previously, the maternal selection indices consisted of more traits, which could have decreased the rate

of genetic progress for any single trait relative to an index consisting of fewer traits. This difference in breeding objective between these two groups of genetic lines is likely responsible for higher AGE SNP effects in the Duroc pigs. Crossbred pigs in The Maschhoff's genetic selection program are not exposed to direct selection pressure. Rather, artificial selection occurs in the three genetic lines that constitute the genetic makeup of the Crossbred population. Mean absolute values for AGE SNP effects in significant associations in the Crossbred pigs (1.80 months) were similar to values reported for the three pure populations, suggesting that selection in the three purebred populations also changes allele frequencies in the Crossbred population at similar rates. However, it must be noted that genotype samples in Crossbred population were collected over a span of approximately 4 years, as opposed to a span of approximately 10 years in the three purebred populations (Table 2.2). We calculated mean yearly change in allele frequency for GPSM significant SNPs in each population, and results for the three purebred populations were similar (0.018, 0.019, and 0.019 for the Duroc, Landrace, and Yorkshire populations, respectively). The mean yearly change in allele frequency for significant SNPs in the Crossbred population was considerably higher than the purebred populations (0.024 vs. 0.018 to 0.019 per year). The ranking of the values of mean yearly allele frequency change for significant SNPs in each population was different than the ranking of mean absolute values for SNP effects. This difference in results is likely due to the adjustment to SNP effects by inclusion of the genomic relationship matrices in the GPSM models, which allows for a more robust estimation of single-SNP selection proxies and more well-calibrated P-values than single-SNP regressions of year on allele frequency.

CONCLUSIONS

We evaluated Generation Proxy Selection Mapping as an analytical method for detecting large and small signatures of artificial selection in a large commercial population of pigs from three purebred populations and one crossbred population. Numerous significant SNPs were detected across the genome in each genetic line, indicating that GPSM is effective in detecting changes in pig genomes due to polygenic selection over relatively short time scales (~4 to 10 years). In addition, simulations proved that GPSM is well-calibrated to distinguish between allele frequency changes over time resulting from genetic drift or artificial selection. Several SNPs were identified as significantly associated with AGE across multiple populations, indicating similar selection objectives, genetic architectures, and causal variants underlying quantitative traits influencing allele frequencies at loci in each population over time. Results from this analysis and future analyses using GPSM could give valuable insight into biological mechanisms underlying and responding to selection on quantitative phenotypes in the commercial swine industry. Lastly, SNPs identified as significantly associated with AGE have the potential to serve as indicators of genomic regions to highlight in the development of genetic prediction models and selection schemes in swine breeding programs.

DECLARATIONS

Ethics approval and consent to participate

Because phenotypic records and tissue samples were collected as part of routine livestock production practices, and were obtained from an existing industry database, ACUC approval was not necessary.

Consent for publication

Not applicable.

Availability of data and materials

Datasets supporting the conclusions of this article are available for non-commercial use via a data use agreement (DUA) with The Maschhoff's, LLC.

Funding

Caleb Grohman was supported by a USDA FFAR Fellows Program grant with matching funds from the Maschhoff's, LLC.

COMPETING INTERESTS

The authors declare that they have no competing interests. Dr. Decker is on the scientific advisory board of Vytelle, LLC.

AUTHORS CONTRIBUTIONS

JED and CJG conceptualized and designed the research. CJG managed data acquisition, storage, and retrieval. CJG estimated variance components and performed association analyses. All authors interpreted results. CJG and JED wrote the initial version of the manuscript, which was edited by all authors. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

We acknowledge research managers of The Maschhoff's, LLC for collection and curation of the genotype data and meta-data. We appreciate the support by The Maschhoff's, LLC in terms of funding and generation of data for continued collaboration. The computation for this work was performed on the high-performance computing infrastructure provided

by Research Computing Support Services and in part by the National Science Foundation under grant number CNS-1429294 at the University of Missouri, Columbia MO.

FIGURES

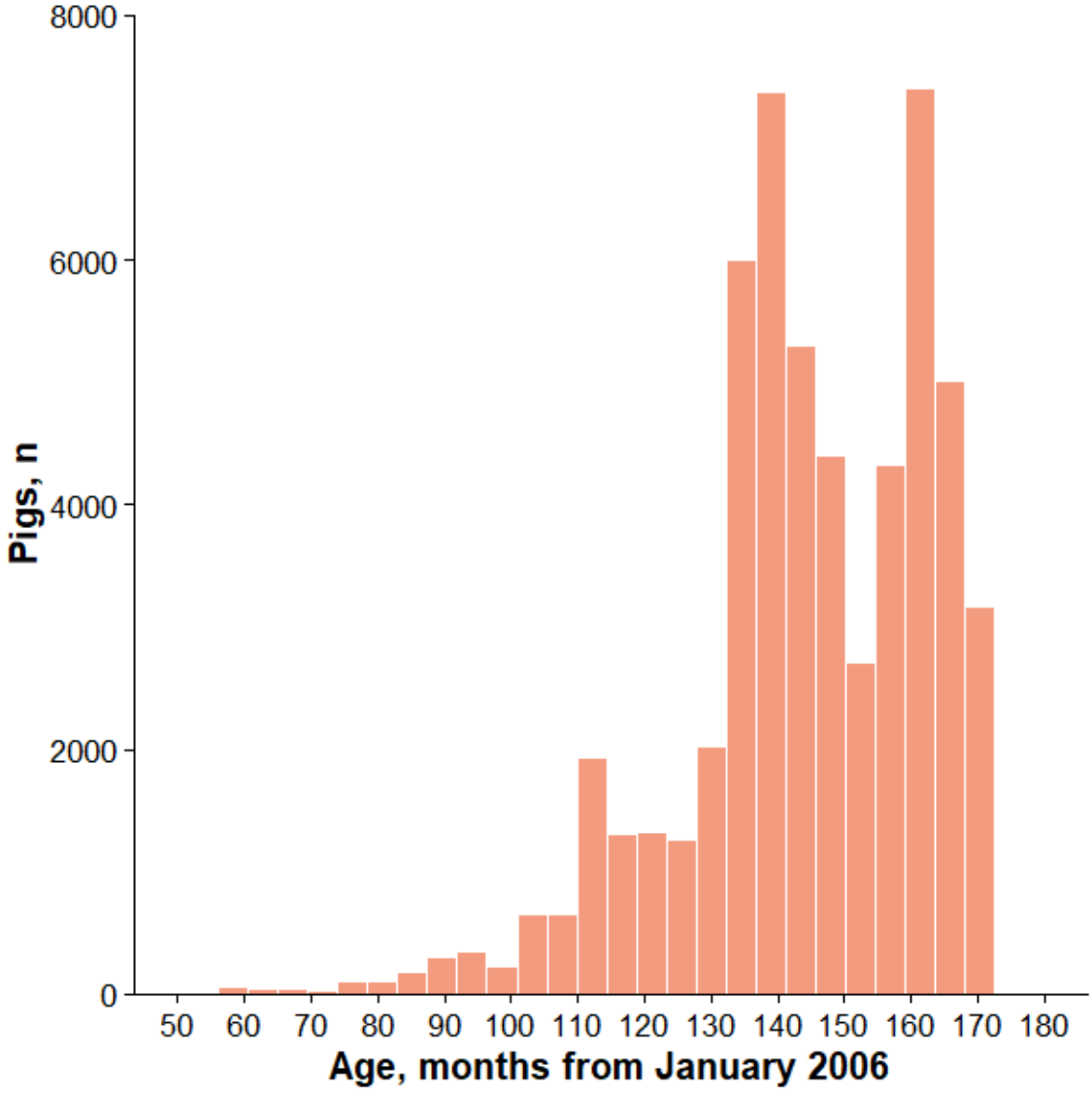


Figure 2.1. Distribution of AGE for all genotyped pigs. For each pig, AGE was calculated as the number of months between each pig's birth month and January 2006. A pig with a negative, zero, or positive AGE was born before January 2006, during January 2006, or after January 2006, respectively.

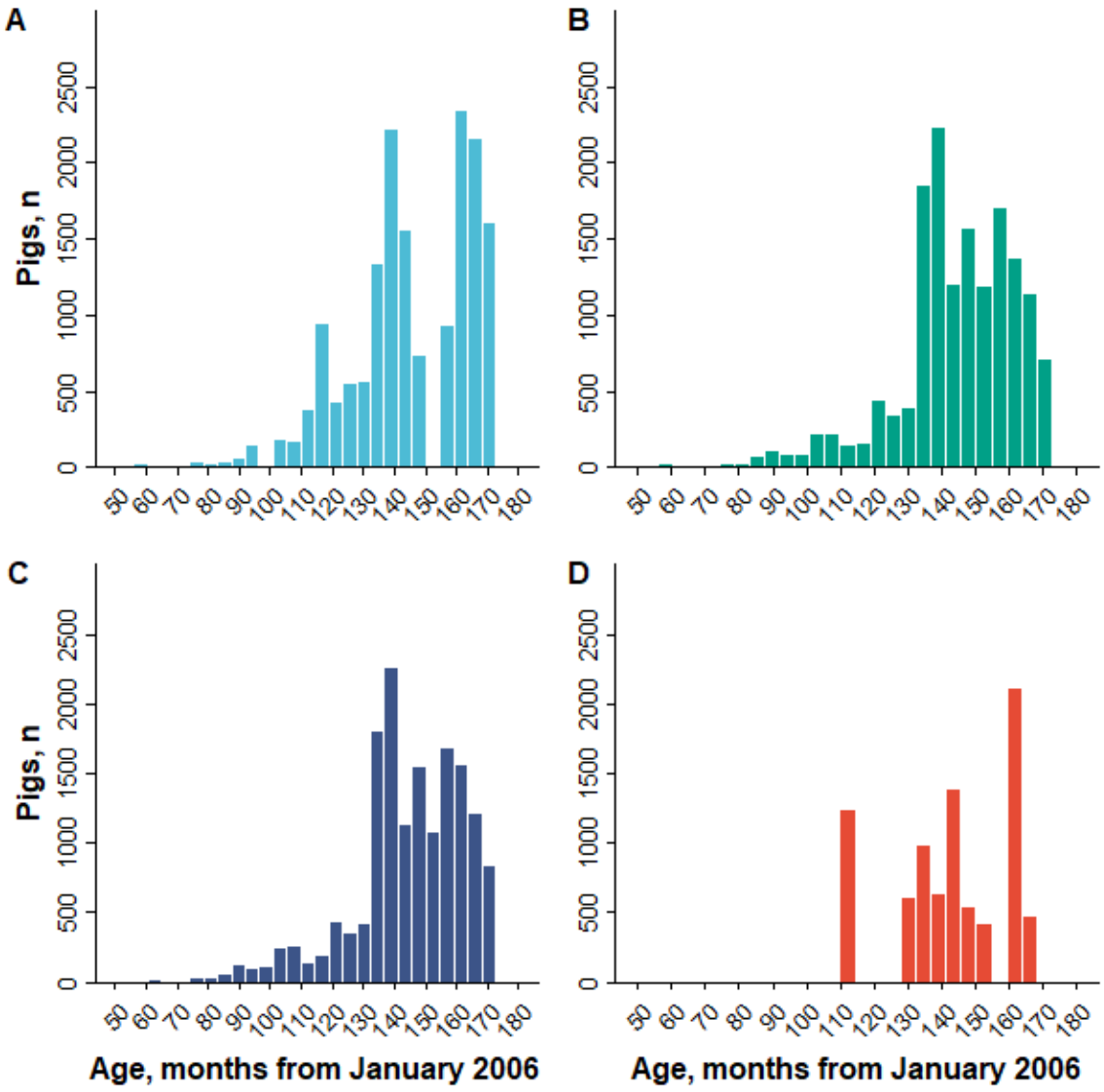


Figure 2.2. Distributions of AGE for genotyped pigs in Duroc, Landrace, Yorkshire, and Crossbred lines. For each Duroc (**A**), Landrace (**B**), Yorkshire (**C**), and Crossbred (**D**) pig, AGE was calculated as the number of months between each pig's birth month and January 2006. For example, a pig with an age of 120 was born in January 2016.

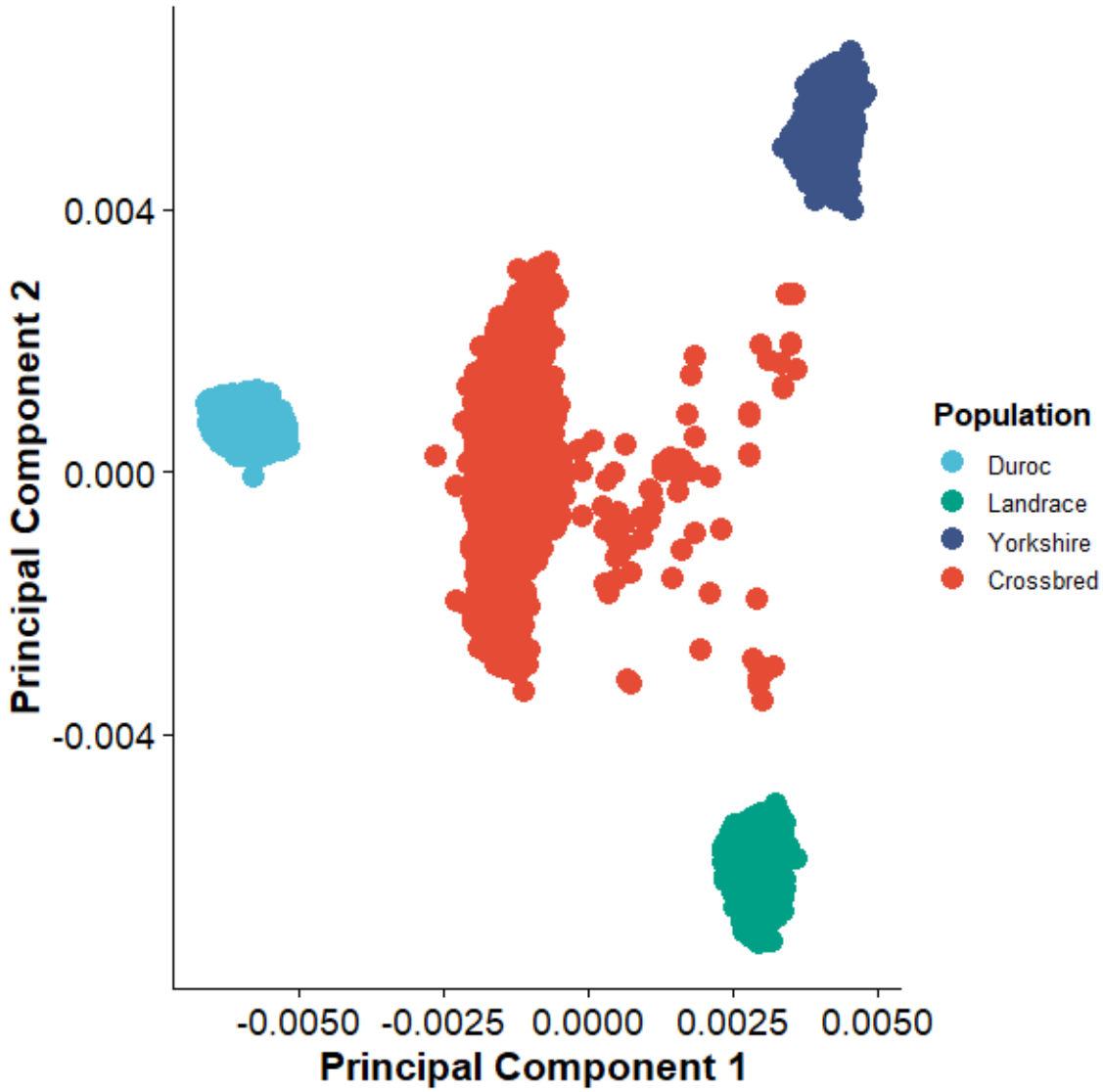


Figure 2.3. Principal components analysis of GRM containing analyzed genotyped pigs. Four clusters appear in the scatterplot. Individual pigs within each colored cluster constitute a population. Populations of pigs located closer in proximity to one another are genetically more similar.

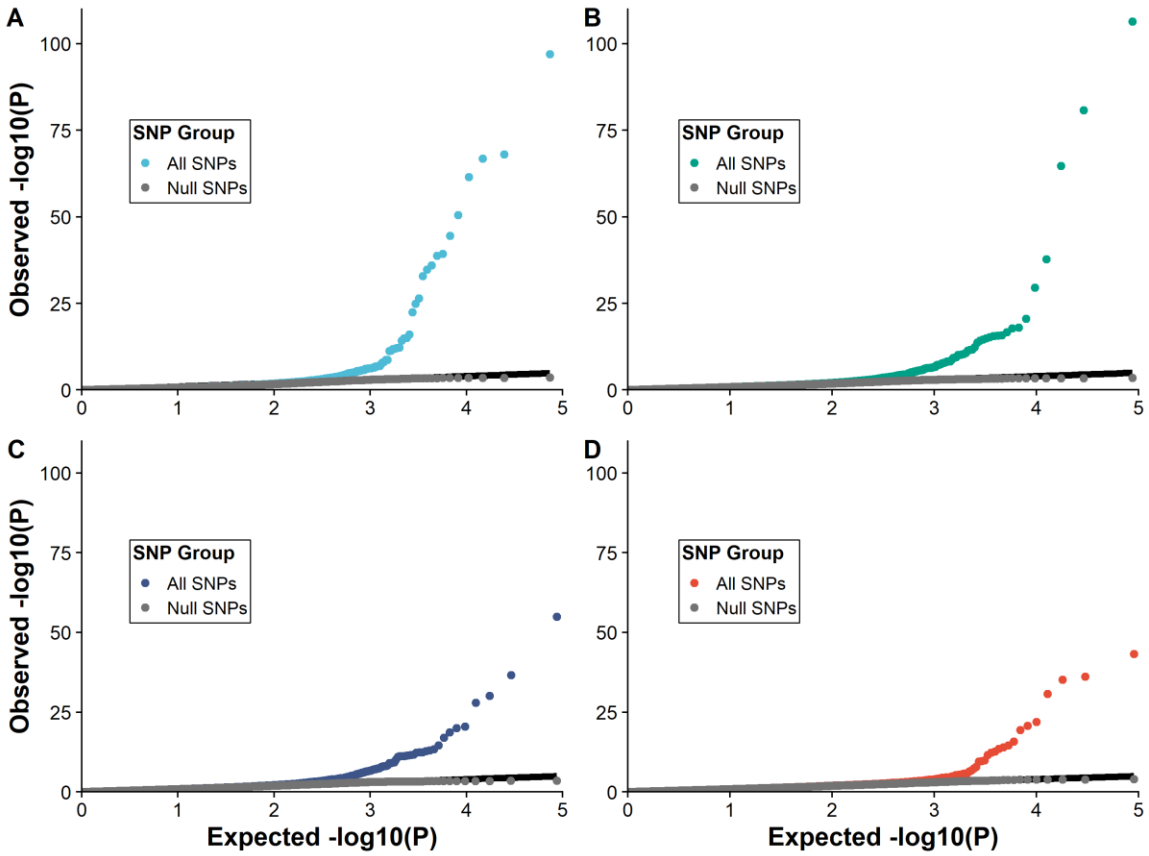


Figure 2.4. Q-Q plots for GPSM P-values from genome-wide association analyses of SNP genotype on AGE. Null SNPs (non-significant) closely followed a uniform distribution, while GPSM significant SNPs deviated from the expected uniform distribution for Duroc (**A**), Landrace (**B**), Yorkshire (**C**), and Crossbred (**D**).

Figure 2.5

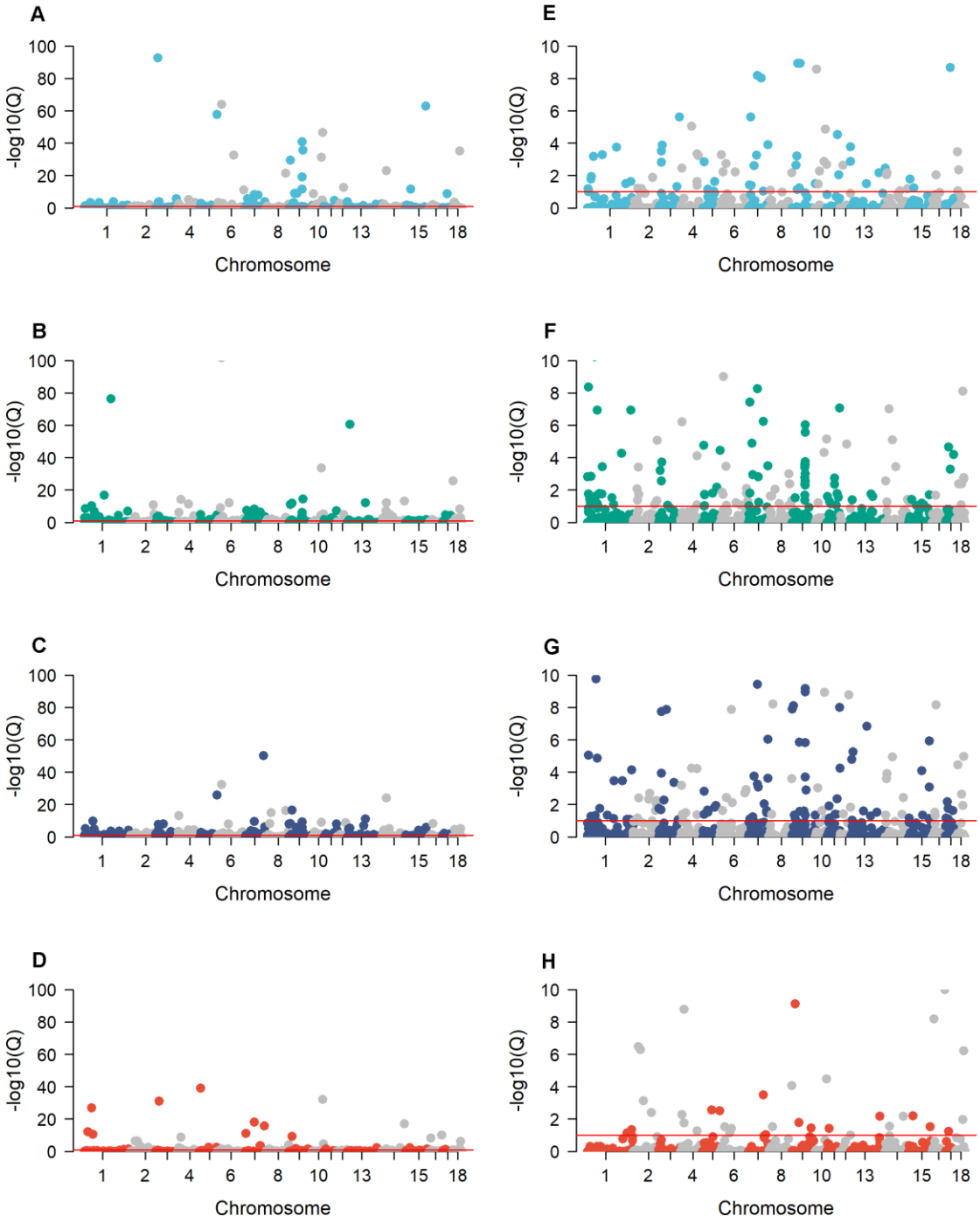


Figure 2.5. Manhattan plots of GPSM Q-values for the association between SNP genotype and AGE. Significant GPSM SNPs were found on each chromosome, and -$\log_{10}$(*Q*-values) and are shown in Manhattan plots with full Y-axes for Duroc (**A**), Landrace (**B**), Yorkshire (**C**), and Crossbred (**D**). Truncated Y-axes from 0 to 10 -$\log_{10}$(*Q*-values) reveal polygenic nature of selection in Duroc (**E**), Landrace (**F**), Yorkshire (**G**), and Crossbred (**H**).

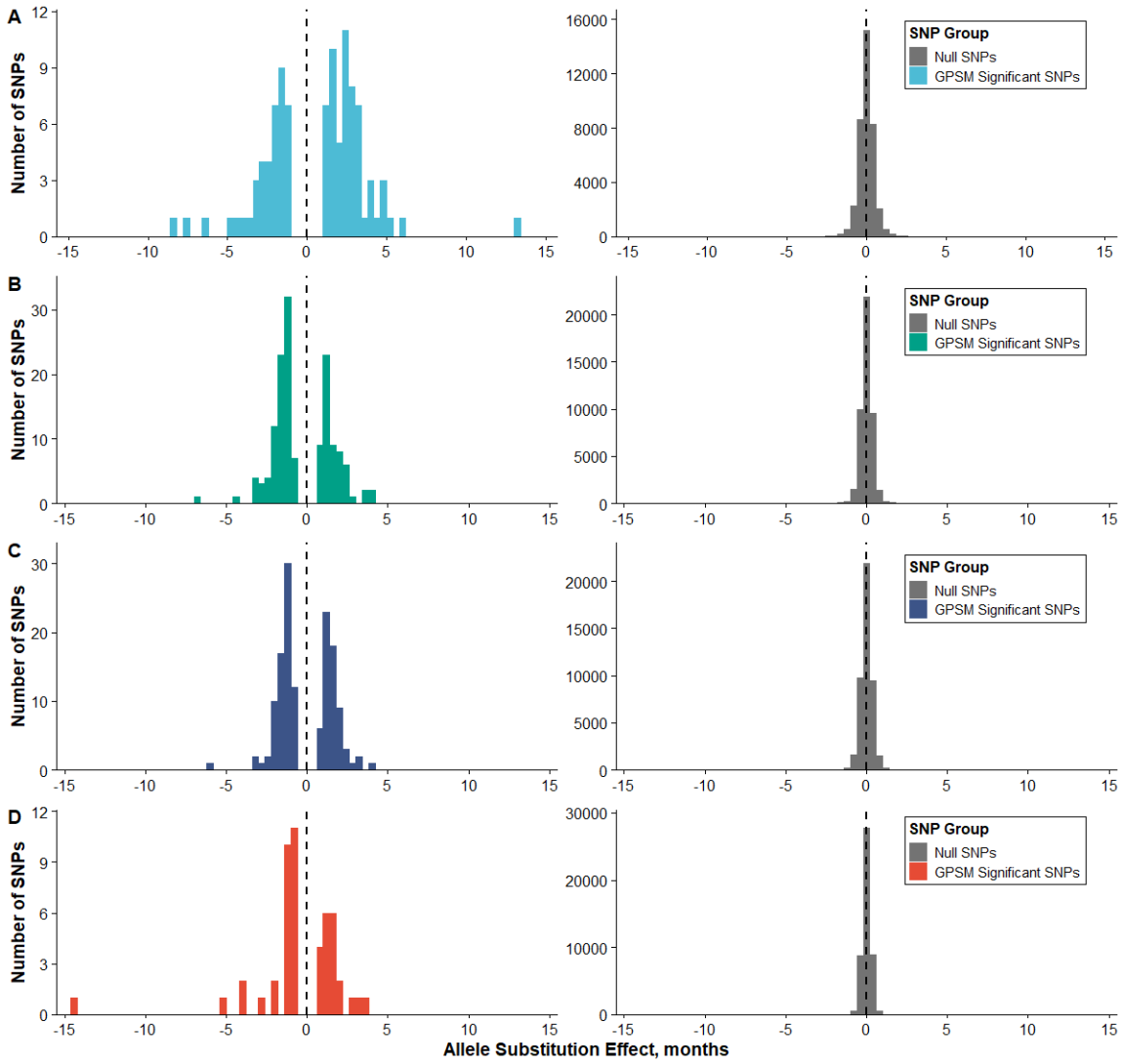


Figure 2.6. Distribution of SNP effects for null and GPSM significant markers. For Duroc (**A**), Landrace (**B**), Yorkshire (**C**), and Crossbred (**D**), null SNPs (non-significant) were normally distributed with a mean near zero, while GPSM significant SNPs followed a bimodal distribution with central values for each peak located above and below zero.

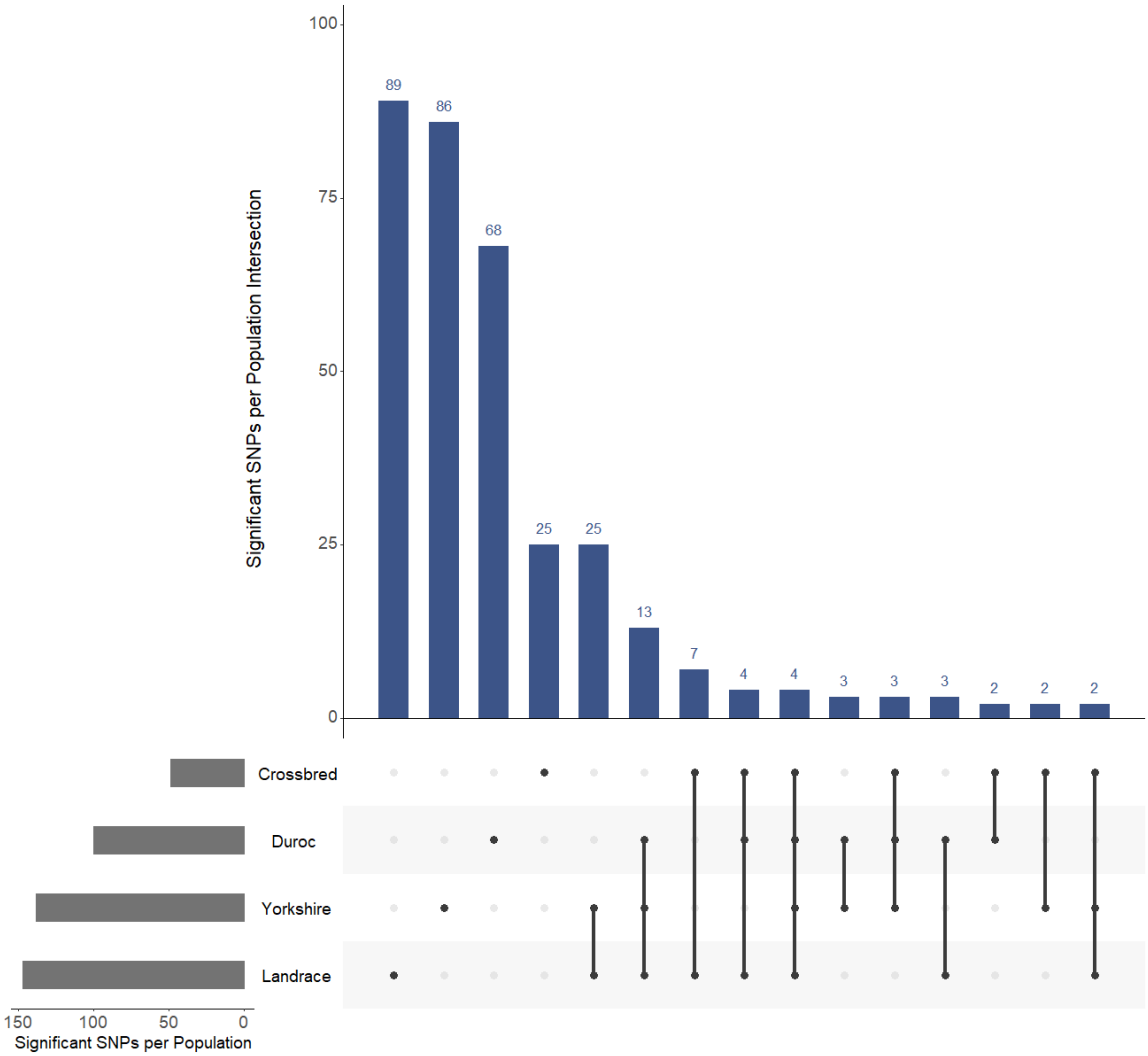


Figure 2.7. UpSet plot depicting the number of GPSM significant SNPs across populations.

Each vertical blue bar shows the number of GPSM significant SNPs unique to a single population (25 to 89 SNPs), unique across two populations (2 to 25 SNPs), unique across three populations (2 to 13 SNPs), or unique across all four populations (4 SNPs). Horizontal gray bars present the number of GPSM significant SNPs in each genetic line.

TABLES

Table 2.1. Summary of pedigree records for all pigs.

Population	Pigs, n	Founders, n[1]	Sires, n	Dams, n	Birth month and year		Generations, n
					Minimum	Maximum	
Duroc	114,038	742	939	6190	March, 1982	September, 2020	14
Landrace	236,385	1778	706	12,856	January, 1993	September, 2020	11
Yorkshire	207,366	730	765	10,749	June, 1980	September, 2020	14
Crossbred	690,193	2025	647	19,627	March, 2015	August, 2020	14

[1]A founder was a pig of generation 0; only the sire side of the pedigree was known for Crossbred pigs

Table 2.2. Summary of pedigree records for all genotyped pigs.

					Birth month and year		
Population	Pigs, n	Founders, n[1]	Sires, n	Dams, n	Minimum	Maximum	Generations, n
Duroc	16,802	17	500	3596	August, 2010	April, 2020	14
Landrace	19,342	82	512	5862	August, 2010	April, 2020	10
Yorkshire	18,368	18	446	5367	January, 2011	April, 2020	14
Crossbred	8532	-	206	4428	March, 2015	September, 2019	13

[1]A founder was a pig of generation 0; only the sire side of the pedigree was known for Crossbred pigs

Table 2.3. Summary of subsets of genotyped pigs and conducted analyses after genotype quality control.

				Analyses[1]		
Subset	Populations	Pigs, n	SNPs, n	Univariate VCE	Bivariate VCE	GWAS
1	Duroc	16,595	38,294	X		X
2	Landrace	15,457	45,085	X		X
3	Yorkshire	15,772	45,027	X		X
4	Crossbred	8447	46,529	X		X
5	Duroc and Landrace	32,066	45,999	X	X	X
6	Duroc and Yorkshire	32,387	46,106	X	X	X
7	Duroc and Crossbred	25,053	46,341	X	X	X
8	Landrace and Yorkshire	31,240	46,253	X	X	X
9	Landrace and Crossbred	23,905	46,440	X	X	X
10	Yorkshire and Crossbred	24,230	46,449	X	X	X
11	Duroc, Landrace and Yorkshire	47,849	46,428	X		X
12	Duroc, Landrace, and Crossbred	40,513	46,415	X		X
13	Duroc, Yorkshire, and Crossbred	40,837	46,424	X		X
14	Landrace, Yorkshire, and Crossbred	39,688	46,458	X		X
15	Duroc, Landrace, Yorkshire, and Crossbred	56,296	46,456	X		X

[1]VCE = variance component estimation; GWAS = genome-wide association study

Table 2.4. Descriptive statistics of AGE by subset.

Subset	Populations	Pigs, n	Mean	SD[1]	Minimum	Maximum
1	Duroc	16,595	144.9	20.44	55	171
2	Landrace	15,457	144.0	18.23	55	171
3	Yorkshire	15,772	144.2	18.46	64	171
4	Crossbred	8447	142.8	17.20	110	164
5	Duroc and Landrace	32,066	144.5	19.41	55	171
6	Duroc and Yorkshire	32,387	144.6	19.50	55	171
7	Duroc and Crossbred	25,053	144.2	19.43	55	171
8	Landrace and Yorkshire	31,240	144.1	18.34	55	171
9	Landrace and Crossbred	23,905	143.6	17.88	55	171
10	Yorkshire and Crossbred	24,230	143.7	18.04	64	171
11	Duroc, Landrace and Yorkshire	47,849	144.4	19.10	55	171
12	Duroc, Landrace, and Crossbred	40,513	144.1	18.98	55	171
13	Duroc, Yorkshire, and Crossbred	40,837	144.2	19.06	55	171
14	Landrace, Yorkshire, and Crossbred	39,688	143.8	18.11	55	171
15	Duroc, Landrace, Yorkshire, and Crossbred	56,296	144.2	18.84	55	171

[1]SD = standard deviation

Table 2.5. Proportion of variation in AGE explained by SNPs for each subset.

Subset	Populations	Pigs, n	SNPs, n	PVE[1]	SE[1]
1	Duroc	16,595	38,294	0.94	0.002
2	Landrace	15,457	45,085	0.87	0.004
3	Yorkshire	15,772	45,027	0.86	0.004
4	Crossbred	8447	46,529	0.94	0.004
5	Duroc and Landrace	32,066	45,999	0.89	0.001
6	Duroc and Yorkshire	32,387	46,106	0.84	0.001
7	Duroc and Crossbred	25,053	46,341	0.91	0.002
8	Landrace and Yorkshire	31,240	46,253	0.89	0.002
9	Landrace and Crossbred	23,905	46,440	0.88	0.003
10	Yorkshire and Crossbred	24,230	46,449	0.87	0.003
11	Duroc, Landrace and Yorkshire	47,849	46,428	0.82	0.001
12	Duroc, Landrace, and Crossbred	40,513	46,415	0.84	0.001
13	Duroc, Yorkshire, and Crossbred	40,837	46,424	0.81	0.001
14	Landrace, Yorkshire, and Crossbred	39,688	46,458	0.87	0.001
15	Duroc, Landrace, Yorkshire, and Crossbred	56,296	46,456	0.84	0.001

[1]PVE = proportion of variation in AGE explained by SNPs (i.e., SNP heritability); SE = standard error

Table 2.6. Proportion of variation in AGE explained by SNPs for each purebred subset using simulated data.

Subset	Populations	Pigs, n	SNPs, n	PVE[1]	SE[1]
1	Duroc	16,595	38,286	0.83	0.005
2	Landrace	15,457	45,090	0.85	0.004
3	Yorkshire	15,772	45,036	0.85	0.005
5	Duroc and Landrace (Method 1)[2]	32,066	46,008	0.90	0.003
5	Duroc and Landrace (Method 2)	32,066	46,008	0.88	0.003
6	Duroc and Yorkshire (Method 1)	32,387	46,098	0.89	0.003
6	Duroc and Yorkshire (Method 2)	32,387	46,098	0.88	0.003
8	Landrace and Yorkshire (Method 1)	31,240	46,260	0.91	0.002
8	Landrace and Yorkshire (Method 2)	31,240	46,260	0.89	0.003
11	Duroc, Landrace, and Yorkshire (Method 1)	47,849	46,422	0.93	0.002
11	Duroc, Landrace, and Yorkshire (Method 2)	47,849	46,422	0.90	0.002

[1]PVE = proportion of variation in AGE explained by SNPs (i.e., SNP heritability); SE = standard error

[2]Method 1 = Genotypes simulated as if populations recently diverged (same founder population); Method 2 = Genotypes simulated as if populations are completely unrelated (different founder populations)

Table 2.7. Genetic correlations for AGE between each pairwise combination of Populations 1 through 4.

Subset	Populations	Pigs, n	SNPs, n	r_G[1]	SE[1]
5	Duroc and Landrace	32,066	45,999	0.64	0.004
6	Duroc and Yorkshire	32,387	46,106	0.67	0.023
7	Duroc and Crossbred	25,053	46,341	0.50	0.018
8	Landrace and Yorkshire	31,240	46,253	0.80	0.017
9	Landrace and Crossbred	23,905	46,440	0.38	0.021
10	Yorkshire and Crossbred	24,230	46,449	0.43	0.020

[1]r_G = genetic correlation; SE = standard error

Table 2.8. Genetic correlations for AGE between each pairwise combination of Populations 1 through 3 using simulated genotype data.

Subgroup	Genetic lines	Pigs, n	SNPs, n	r_G[1]	SE[1]
5	Duroc and Landrace (Method 1)[2]	32,066	46,008	-0.03	0.028
5	Duroc and Landrace (Method 2)	32,066	46,008	-0.02	0.058
6	Duroc and Yorkshire (Method 1)	32,387	46,098	0.06	0.029
6	Duroc and Yorkshire (Method 2)	32,387	46,098	0.03	0.057
8	Landrace and Yorkshire (Method 1)	31,240	46,260	-0.02	0.028
8	Landrace and Yorkshire (Method 2)	31,240	46,260	-0.01	0.056

[1]r_G = genetic correlation; SE = standard error

[2]Method 1 = Genotypes simulated as if populations recently diverged (same founder population); Method 2 = Genotypes simulated as if populations diverged several years ago (different founder populations)

Table 2.9. Number of SNPs significantly associated with AGE for each subset.

Subset	Populations	Pigs, n	SNPs, n	Significant SNPs, n[1]
1	Duroc	16,595	38,294	100
2	Landrace	15,457	45,085	147
3	Yorkshire	15,772	45,027	138
4	Crossbred	8447	46,529	49
5	Duroc and Landrace	32,066	45,999	371
6	Duroc and Yorkshire	32,387	46,106	527
7	Duroc and Crossbred	25,053	46,341	148
8	Landrace and Yorkshire	31,240	46,253	177
9	Landrace and Crossbred	23,905	46,440	172
10	Yorkshire and Crossbred	24,230	46,449	182
11	Duroc, Landrace and Yorkshire	47,849	46,428	702
12	Duroc, Landrace, and Crossbred	40,513	46,415	533
13	Duroc, Yorkshire, and Crossbred	40,837	46,424	609
14	Landrace, Yorkshire, and Crossbred	39,688	46,458	274
15	Duroc, Landrace, Yorkshire, and Crossbred	56,296	46,456	854

[1] $Q < 0.10$

Table 2.10. Number of SNPs significantly associated with AGE for each subset using randomly simulated genotype data.

Subset	Populations	Pigs, n	SNPs, n	Significant SNPs, n[1]	Error Rate, %[3]
1	Duroc	16,595	38,286	1	0.0026
2	Landrace	15,457	45,090	0	0.0000
3	Yorkshire	15,772	45,036	2	0.0044
5	Duroc and Landrace (Method 1)[2]	32,066	46,008	0	0.0000
5	Duroc and Landrace (Method 2)	32,066	46,008	4	0.0087
6	Duroc and Yorkshire (Method 1)	32,387	46,098	0	0.0000
6	Duroc and Yorkshire (Method 2)	32,387	46,098	7	0.0152
8	Landrace and Yorkshire (Method 1)	31,240	46,260	1	0.0022
8	Landrace and Yorkshire (Method 2)	31,240	46,260	3	0.0065
11	Duroc, Landrace, and Yorkshire (Method 1)	47,849	46,422	0	0.0000
11	Duroc, Landrace, and Yorkshire (Method 2)	47,849	46,422	6	0.0129

[1]$Q < 0.10$

[2]Method 1 = Genotypes simulated as if populations recently diverged (same founder population); Method 2 = Genotypes simulated as if populations are completely unrelated (different founder populations)

[3]Error Rate = (Significant SNPs, n/SNPs, n) × 100

Table 2.11. Ten SNPs significantly associated with AGE with the largest absolute values for SNP effects within the Duroc, Landrace, Yorkshire, and Crossbred populations.[1]

Subset	Populations	SNP Identifier	MAF	SNP Effect	SE	*Q*-value	Number of subsets where significant
1	Duroc	3420	0.24	13.32	0.310	1.49E-102	3
		41017	0.02	8.21	0.648	3.54E-33	1
		30819	0.07	7.71	0.368	2.05E-93	2
		35689	0.09	6.49	0.390	2.34E-58	3
		39502	0.47	6.19	0.353	1.24E-64	3
		18513	0.02	5.16	0.643	2.28E-12	1
		30855	0.13	4.87	1.025	1.56E-03	1
		49794	0.38	4.79	0.340	1.56E-41	3
		31005	0.11	4.66	0.880	1.31E-04	1
		6055	0.08	4.60	0.370	5.40E-32	3
2	Landrace	39502	0.30	6.64	0.302	1.76E-102	3
		14465	0.02	4.40	0.518	9.00E-14	3
		1063	0.46	3.95	0.207	3.69E-77	1
		22747	0.31	3.81	0.333	2.43E-26	1
		10745	0.42	3.52	0.206	2.66E-61	2
		38925	0.04	3.51	0.840	1.55E-02	1
		6055	0.18	3.08	0.238	2.03E-34	3
		485	0.23	3.08	0.325	1.71E-17	2
		39264	0.28	3.03	0.418	9.58E-10	1
		8174	0.29	3.00	0.456	8.35E-08	2
3	Yorkshire	45804	0.02	6.00	0.664	1.14E-15	1
		39502	0.36	4.03	0.316	4.80E-33	3
		19756	0.17	3.40	0.486	7.20E-09	1
		35689	0.28	3.28	0.284	1.24E-26	3
		9883	0.39	3.05	0.385	1.20E-11	2
		8174	0.14	3.03	0.437	1.02E-08	2
		41553	0.43	2.96	0.188	5.05E-51	1
		30156	0.04	2.87	0.790	9.16E-02	1
		34751	0.05	2.52	0.461	5.84E-05	3
		34530	0.18	2.38	0.277	6.05E-14	2
4	Crossbred	30819	0.07	14.39	0.372	1.81E-102	2
		14465	0.02	5.10	0.536	9.90E-18	3
		3834	0.04	4.01	0.344	1.58E-27	2
		36398	0.18	4.00	0.288	1.25E-39	2
		31018	0.48	3.52	0.281	7.71E-32	1
		6187	0.40	3.06	0.241	1.00E-32	3
		3420	0.43	2.99	0.271	2.72E-24	3
		43184	0.08	2.77	0.282	7.68E-19	4
		33882	0.07	2.09	0.290	1.63E-09	3
		4012	0.13	2.00	0.258	4.15E-11	1

[1]MAF = minor allele frequency; SE = standard error

SUPPLEMENTARY TABLES

Supplementary Table 2.1. Proportion of variation in AGE explained by SNPs for each purebred subset using five replications of randomly simulated genotype data.

				PVE[1]						
Subset	Population	Pigs, n	SNPs, n	Replication 1	Replication 2	Replication 3	Replication 4	Replication 5	Mean	SEM[2]
1	Duroc	16,595	38,286	0.83	0.84	0.82	0.83	0.83	0.83	0.004
2	Landrace	15,457	45,090	0.85	0.87	0.86	0.86	0.86	0.86	0.003
3	Yorkshire	15,772	45,036	0.84	0.84	0.85	0.84	0.84	0.84	0.002

[1]PVE = proportion of variation in AGE explained by SNPs
[2]SEM = standard error of the mean PVE across replicates

Supplementary Table 2.2. Number of SNPs significantly associated with AGE for each subset using five replicates of randomly simulated genotype data.

				Replicate ID						
Subset	Population	Pigs, n	SNPs, n	1	2	3	4	5	Mean	SEM
False positives, n										
1	Duroc	16595	38286	0	0	0	3	3	1.20	0.735
2	Landrace	15457	45090	1	1	0	0	3	1.00	0.548
3	Yorkshire	15772	45036	0	0	1	0	0	0.20	0.200
Error rate, %										
1	Duroc	16595	38286	0.0000	0.0000	0.0000	0.0078	0.0078	0.0031	0.00192
2	Landrace	15457	45090	0.0022	0.0022	0.0000	0.0000	0.0067	0.0022	0.00121
3	Yorkshire	15772	45036	0.0000	0.0000	0.0022	0.0000	0.0000	0.0004	0.00044

Supplementary Table 2.3 is too large for print, but can be found in the following folder:

https://drive.google.com/drive/folders/1fv_iCxmQkEmX7y5PqaIWJOR-_zcWNWx_?usp=sharing

CHAPTER 3. TECHNICAL NOTE: A NOVEL ALGORITHM TO IDENTIFY MORTALITY EPISODES IN COMMERCIAL WEAN-TO-FINISH PIG COHORTS

Caleb J. Grohmann[†], Caleb M. Shull[‡], and Jared E. Decker [†§ψ*]

[†]Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA

[‡]The Maschhoff's, LLC, Carlyle, IL 62231, USA

[§]Genetics Area Program, University of Missouri, Columbia, MO 65211, USA

[ψ]Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA

*Corresponding author

LAY SUMMARY

The number of dead or euthanized pigs per day in commercial wean-to-finish pig barns can be hard to accurately predict. As wean-to-finish mortality in pigs has realized increases recently, methods to forecast dead and euthanized pigs are critical to mitigate economic loss and improve pig health, productivity, and well-being. We introduce the concept of mortality episodes, which are defined as sequences of days where the average daily mortality is significantly higher than the population baseline at a given age for a unique cohort of pigs reared in a single room on a single farm. We developed a novel method to identify mortality episodes in wean-to-finish pig barns, which revealed periods of time in which mortality was the most intense and sustained. Results from this analysis can be used in predictive models as the dependent variable, which will allow producers and researchers to proactively anticipate the onset of mortality episodes with the goal of reducing overall mortality rate in wean-to-finish pig production in real-time.

TEASER TEXT

Mortality episodes in commercial wean-to-finish pig barns are defined as sequences of days where the average daily mortality is significantly higher than the population baseline at a given age for a unique cohort of pigs reared in a single room on a single farm. We propose an algorithm to identify mortality episodes, which can serve as the basis for models that predict the onset of mortality episodes in real-time.

ABSTRACT

Daily mortality in commercial wean-to-finish pig barns is highly variable, influenced by several factors, and is hard to predict using current statistical methods. We propose the concept of mortality episodes, which are defined as sequences of days where the average daily mortality is significantly higher than the population baseline at a given age for a unique cohort of pigs reared in a single room on a single farm. Mortality episodes can be described by three main attributes: time, duration, and magnitude. Here, we developed and tested a novel algorithm based on penalized smoothing spline regression to classify mortality episodes in commercial wean-to-finish pig barns. To initialize the algorithm, 5 parameters (PADDING = number of zero values appended before and after the mortality time series, LAMBDA = spline penalization parameter, MAGNITUDE = classification threshold for average number of dead pigs per mortality episode day, DURATION = classification threshold for maximum number of days in an episode, and PROPORTION = classification threshold for the proportion of days in an episode where at least 1 dead pig was observed) are selected that influence the behavior of the underlying regression equation and mortality classifications. Using the selected parameters, penalized smoothing spline regression equations are fit to a mortality time series [independent variable = days post-placement (**DAYS**); dependent variable = daily mortality count or rate (**TMORT**)] from a unique cohort of pigs. For each cohort, DAYS and predicted TMORT is calculated and used to identify changepoints [i.e., peaks (negative instantaneous slope at $DAYS_t$ and positive instantaneous slope at $DAYS_{t-1}$) and valleys (positive instantaneous slope at $DAYS_t$ and negative instantaneous slope at $DAYS_{t-1}$)]. Within each mortality sequence, the maximum and minimum instantaneous

slope were considered the start and end of the mortality episode, respectively. A grid search experiment across possible parameter combinations was performed and tested on mortality time series data from 10,906 daily mortality counts across 82 unique cohorts of pigs. Results from the grid search experiment showed that as the padding and lambda parameters increased, the number of found dead and euthanized pigs per classified mortality episode day decreased, the classified mortality episode duration increased, and the number of classified mortality episodes per cohort decreased. Other parameters, while still important, had less of an effect on the above statistics than the padding and lambda parameters. Selected optimal parameters yielded precise start, peak, and end of mortality episode classifications upon visual inspection of mortality time series from each unique cohort. Mortality episode classifications from the proposed algorithms can serve as the dependent variable in future prediction analyses to proactively predict the onset of mortality episodes in real-time.

INTRODUCTION

Mortality patterns in United States commercial wean-to-finish pig barns are highly variable and influenced by several factors. For example, the number of dead pigs observed daily across the growth period at a typical wean-to-finish pig farm is generally zero-inflated and highly right-skewed (Varona and Sorensen, 2010). Therefore, analysis of daily mortality counts, rates, and occurrence requires complicated statistical models that account for fixed and random effects and non-normally distributed dependent variables, such as binomial, Poisson, negative-binomial, and zero-inflated mixed regression models (Varona and Sorensen, 2010). Due to the described distributional nature and the randomness of mortality data, most of the variation in these variables can be attributed to residual error (Dufrasne et al., 2014; Su et al., 2022). Thus, predicting singular values for future daily mortality is impractical using current statistical methods.

Here, we define mortality "episodes" in commercial wean-to-finish pig barns as sequences of days where the average daily mortality is significantly higher than the population baseline at a given age for a unique cohort of pigs reared in a single room on a single farm. Mortality episodes have three main attributes: 1) time, defined as the day post-weaning the mortality episode starts, 2) duration, defined as the total number of consecutive days in the episode, and 3) magnitude, defined as the average number of dead pigs observed per day for the duration of the episode (Mehling et al., 2019). The statistical analysis of mortality episodes provides advantages over the analysis of daily mortality counts and rates. First, the identification of mortality episodes highlights the periods when mortality risk is the most severe and removes random noise from a mortality time series. Second, by characterizing common patterns of mortality episodes,

advanced statistical and machine learning models can be trained to predict the start of periods of high mortality risk as opposed to singular days of abnormally high mortality, which are subject to elevated random noise. To elucidate mortality episodes, a method to "smooth" the daily mortality data is required, which removes random noise and reveals the general mortality curve for each cohort of wean-to-finish pigs. In addition, identification of the start and end of each mortality episode is necessary. Smoothing methods, such as moving averages, have long been used to reduce noise in time-series data across a wide range of distributions (Macaulay, 1931; Raudys et al., 2013). However, the number of time periods used to calculate the moving average is hard to identify and is subject to considerable variation across pig cohorts; thus, the resulting curve is rarely perfectly "smooth" unless several data points are utilized to calculate each value, which reduces the amount of useful information remaining in the smoothed curve.

Smoothing splines provide an interesting opportunity to remedy many of the disadvantages of moving average smoothing techniques. Splines are generally used in statistics to mathematically reproduce flexible shapes (Perperoglou et al., 2019), much like mortality curves observed in wean-to-finish pig barns (Mehling et al., 2019). Spline functions are derived that minimize the residual sums of squares (**RSS**) between two variables (i.e., days post-weaning and daily mortality counts or rates), which yields a "best-fitting" smoothed curve. Therefore, we propose a novel algorithm based on smoothing spline regression to identify the start, peak, and end of mortality episodes in unique cohorts of commercial wean-to-finish pigs.

MATERIALS AND METHODS

Regression and Smoothing Splines

As previously stated, splines are used in statistical modeling to fit regression equations between variables that possess a nonlinear relationship. Knots, defined as k, a key parameter estimated in a regression spline equation, are placed at points within the independent variable series where additional flexibility in the equation is required. Within a sequence of knots, unique polynomial functions, defined as splines, are fit in a piece-wise manner to represent the nonlinear relationship between the range of independent and dependent variable values between two consecutive knots. Typically, polynomials of order 3 are used for each spline as they result in curves that are generally "perfectly smooth" to the human eye; however, linear (polynomial order 1) or quintic (polynomial order 5) polynomials are sometimes used depending on the required flexibility between successive knots (Perperoglou et al., 2019).

Polynomial Regression Splines. Figure 3.1 shows the relationship between days postweaning and the number of found dead and euthanized pigs in a selected example cohort of pigs from wean-to-finish commercial production. In the case of a time-series of mortality counts in a wean-to-finish barn, the relationship between day postweaning and daily mortality count is generally nonlinear. In addition, this relationship is rarely standard across an entire cohort of pigs from wean-to-finish. A simple regression model of the form $y_i = \beta_0 + \beta_1 x_i + \epsilon$ for $i = 1, 2, \dots, n$ gives a poor estimation of the wean-to-finish mortality time-series curve ("Simple Linear Regression" model in Figure 3.1). This model can be extended to a cubic polynomial regression model of the form $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon$ for $i = 1, 2, \dots, n$ to better fit the nonlinear relationship;

however the cubic equation assumes the same regression coefficients across the range of data ("Cubic Polynomial" regression model in Figure 3.1). Thus, the main advantage of a regression spline model over the simple or cubic polynomial regression model is the estimation of $k + 1$ polynomials of order 3 (i.e., splines) across the entire mortality series (Perperoglou et al., 2019). Consider a regression spline equation with 7 equally spaced knots ($k_1, k_2, \dots, k_7$) across a range of days postweaning values (24.75, 49.5, 75.25, 99, 123.75, 148.5, and 173.25 in Figure 3.1). The regression spline equation would then assume the following form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 (x_i - k_1)^3 + \beta_5 (x_i - k_2)^3 + \beta_6 (x_i - k_3)^3 + \beta_6 (x_i - k_4)^3 + \beta_7 (x_i - k_5)^3 + \beta_8 (x_i - k_6)^3 + \beta_9 (x_i - k_7)^3 + \epsilon$$

for $i = 1, 2, \dots, n$. The resulting regression spline has a total of $k + 4$ estimated degrees of freedom (4 parameters in each inter-knot sequence knots minus 3 constraints for each knot). In regression spline equations, the flexibility of the model is a direct function of the number of selected knots (e.g., more knots produce a more flexible equation). Because of the added complexity and degrees of freedom of the regression spline equation, the nonlinear relationship between days postweaning and a mortality count series can be modeled more precisely (Figure 3.1) (Perperoglou et al., 2019).

Penalized Smoothing Splines. In a simple regression spline equation ("Regression Spline" in Figure 3.1), the amount and spacing of knots must be explicitly chosen before derivation of the equation. To overcome the problem of knot selection, penalized smoothing splines can be utilized in regression equations. This method assumes a that the nonlinear relationship between two variables can be modeled by a large set of equidistant knots automatically selected by the penalization function, with the maximum number of

possible knots equal to the number of observations in the series, which provides a high degree of flexibility to the equation (Perperoglou et al., 2019). To avoid overfitting the data, a penalty parameter known as λ or "lambda" is introduced to "penalize" the polynomial regression coefficients between each knot. Lambda can assume values from 0 to infinity, with 0 being no penalty (i.e., high flexibility) and infinity being a linear function (i.e., no flexibility). Then, the tuning of parameters to produce the smoothing spline regression equations becomes less complex, with lambda being the only parameter left to optimize, as opposed to the explicit selection of knots (Perperoglou et al., 2019). In this study, we extend the use of penalized smoothing spline equations to identification of mortality episodes in commercial wean-to-finish barns. Therefore, the goal of the proposed algorithm is not to fit the optimal curve (i.e., minimize residual sums of squares) between days postweaning and mortality counts, but to identify periods of acute, intense mortality that spans sequences typically less than 14 to 21 days. Ultimately, we require a smoothing function that allows for several observed peaks across the mortality time-series, which represent potential mortality episodes, and this is best accomplished by utilization of penalized smoothing spline equations ("Penalized Smoothing Spline" in Figure 3.1).

Algorithm to Identify Mortality Episodes

All steps in the proposed algorithm for identification of the start, peak, and end of mortality episodes in commercial wean-to-finish pig barns were implemented using the 'dplyr' package (Wickham et al., 2023) and 'base' and 'stats' packages of R v4.3.0 (R Core Team, 2023). Each step is described in detail below:

Applied to Entire Dataset:

Step 1. Split dataset into multiple smaller datasets by group of pigs within room within site.

Applied to each Split Dataset(s) from Step 1:

Step 2. Add leading and lagging zero-value observations to the Y (daily mortality rate or count; referred to here as TMORT) variable, with X (time series; referred to here as DAYS) set sequentially according to the amount of leading and lagging observations. The number of leading and lagging zeros can vary, but more than 1000 is not recommended, in general. This parameter is denoted as PADDING.

Step 3. Estimate a penalized smoothing spline regression equation between DAYS and TMORT variables from Step 2 using the **smooth.spline** function of the 'stats' package in R. Select desired lambda value (i.e., smoothing parameter; parameter name: LAMBDA) for use in the **smooth.spline** function. Values of LAMBDA that are closer to zero are less restrictive and will produce a more flexible regression equation.

Step 4. Extract predicted values from estimated smoothing spline regression curve, defined as $\hat{y}_t$. Calculate instantaneous slope, defined as $\frac{\Delta y}{\Delta x}$, for each observation using the following formula:

$$\frac{\Delta y}{\Delta x} = \frac{\hat{y}_t - \hat{y}_{t-1}}{DAYS_t - DAYS_{t-1}}$$

Step 5. Identify changepoints, which are days where the sign of the instantaneous slope is different from the previous day. Definitions of the two possible changepoints are described below:

1) Valley: Positive instantaneous slope on $DAYS_t$ and negative instantaneous slope on $DAYS_{t-1}$.

2) Peak: Negative instantaneous slope on $DAYS_t$ and positive instantaneous slope on $DAYS_{t-1}$.

Step 6. Identify mortality sequences, which are all days between each valley. For example, if a series of days has 3 valleys (A, B, and C), there would be 2 mortality sequences in the series (mortality sequence 1 = all days between valley A and B and mortality sequence 2 = all days between valley B and C).

Step 7. Within each mortality sequence, find the days with the maximum (i.e., fastest increasing) and minimum (i.e., fastest decreasing) instantaneous slopes. Consider the maximum and minimum instantaneous slopes in each mortality sequence the start and end of the given mortality sequence, respectively. Then, all days between the start and end day of the mortality sequence are considered a "first-pass" mortality episode.

Final Classification of Mortality Episodes:

Step 8. Declassify first-pass mortality episodes that fail to meet the following criteria or minimum parameter values:

1) CRITERIA: The mortality episode contains a start, peak, and end.
2) MAGNITUDE PARAMETER: The average number of found dead or euthanized pigs per day during the mortality episode must be greater than or equal to MAGNITUDE.
3) PROPORTION PARAMETER: The proportion of days during the mortality episode with at least 1 found dead or euthanized pig must be greater than or equal to PROPORTION.

4) DURATION PARAMETER: The number of days in the mortality episode must not exceed DURATION.

After the above algorithm is applied to a given dataset, the output contains a modified version of the original dataset that details the classification of each day. Two key variables are appended to the original dataset: 1) Day Classification, which contains four possible values (Normal Day, Start of Episode, Peak of Episode, and End of Episode) and 2) Episode Classification, which contains two possible values (0 = a day that was not during a mortality episode and 1 = a day that was during a mortality episode).

Application to Wean-to-Finish Mortality Data

Animals and Facilities. Animal Care and Use Committee approval was not needed because the data was obtained from a preexisting commercial database. The data for all analyses were collected from 98 different groups of pigs grown at six different wean-to-finish sites owned and operated by The Maschhoff's, LLC (Carlyle, Illinois, USA) between December 2020 to October 2023. A group of pigs was considered a unique cohort of pigs allotted to a single room within a single site. Each of the six wean-to-finish sites contained two rooms of pigs under study, which totaled twelve rooms across the entire trial. Three of the six sites were in Iowa and were considered commercial wean-to-finish sites, where only standard pig feeding and growth occurred. Rooms in each commercial site were 50 feet × 193 feet in dimension and contained fourteen separate pens. The other three sites in the current trial were in central and southern Illinois and were considered commercial research wean-to-finish sites. In these research sites, groups of weaning pigs were routinely allotted and evaluated for growth performance and

mortality measures until the end of the finishing phase as a part of nutrition, health, and genetics trials relevant to commercial pig production. However, these trials were expected to have minimal impact on data collected in the current study. Rooms in each of the research sites were 50 feet × 300 feet in dimension and contained 60 to 90 pens.

Pigs in each group were sired by an AcuFast™ (Saskatoon, Saskatchewan, Canada; formerly Acuity™, Carlyle, Illinois, USA) Duroc terminal sire bred to a Yorkshire × Landrace maternal dam, mixed sex, and weaned at twenty to twenty-one days of age. All sites in this trial were managed as "all-in, all-out", and each room was populated within seven days to reduce health concerns between and within groups of pigs. During the nursery period, which was defined as week zero post-weaning until week seven to twelve post-weaning, pigs were "double-stocked" to approximately three-square feet per pig, which is standard protocol within The Maschhoff's system. "Thin-down" occurred between week seven and twelve post-weaning for all groups, which is defined as the process of removing half of the pigs from each pen to be raised in a different location for the duration of the growing-finishing period. Any pigs removed during the thin-down process were subsequently removed from the present study. After thin-down, all pigs were stocked to a density of six square feet per pig. Finishing pigs, upon reaching a fixed market weight of approximately 285 pounds, were sent to harvest over a period of approximately 6 to 8 weeks. Animal housing, feeding, handling, and veterinary care were under the supervision of The Maschhoff's management personnel. All rooms had fully slatted floors, deep-pit manure handling, mechanically controlled ventilation, and automated feeding and bowl waterers. Pigs were provided *ad libitum*

access to feed and water from weaning to harvest in a wet-dry feeding system and were fed standard commercial corn-soybean pig diets.

All pigs received standard vaccination and medications that followed The Maschhoff's standard protocol, which was detailed by Krahn as follows: *Mycoplasma hyopneumoniae* vaccine (Fostera® Gold PCV MH, Zoetis, Kalamazoo, MI, USA; Circumvent® PCV-M G2, Merck Animal Health, Summit, NJ, USA or Ingelvac MycoFlex®, Boehringer Ingelheim Vetmedica Inc, St. Joseph, MO, USA) at processing (3 to 5 days of age), and at 2-weeks post-weaning, porcine reproductive respiratory syndrome virus modified-live virus vaccine (Ingelvac PRRS® MLV, Boehringer Ingelheim Vetmedica Inc, St. Joseph, MO, USA) at 2-weeks post-weaning, and porcine circovirus type 2 (PCV2) killed vaccine (Fostera Gold PCV® MH, Zoetis, Kalamazoo, MI, USA; Circumvent® PCV-M G2, Merck Animal Health, Summit, NJ, USA or Ingelvac CircoFlex®, Boehringer Ingelheim Vetmedica Inc, St. Joseph, MO, USA) vaccine at 3-weeks post-weaning (Krahn, 2018). Feed medication protocol followed The Maschhoff's standard protocols and were kept consistent between all groups of pigs. All water and injectable antimicrobial treatments and interventions performed were part of the routine care administered to animals by their caretakers.

Data Collection and Preparation. Age and mortality measurements for each room within site were collected by barn and research personnel. Data were summarized by The Maschhoff's research personnel into daily measurements and communicated to researchers at the University of Missouri (Columbia, Missouri, USA) for further summary, calculation, and analysis. The site, room within site, cohort within room within site, calendar date, number of days post placement (**DAYS**), and the total number of pigs

(**INV**) was provided for each daily room measurement. Days post placement was defined as the number of days since the start of allotment. For example, a value of "0" for days post placement would indicate the first day a new group of pigs were moved into a room, while a value of "100" would indicate the 101st day on feed for a given group of pigs in a room. The number of pigs found dead (**DEAD**) and euthanized (**EUTH**) was recorded by barn personnel at the end of each workday. Total mortality was calculated as the sum of the daily number of found dead and euthanized pigs (**TMORT**). After data aggregation, 13,198 daily observations across 98 cohorts of pigs were kept for further analysis. There were no missing observations for the days post placement and inventory variables. However, 290 daily measurements were missing across each of the DEAD, EUTH, and TMORT variables (i.e., the same 290 days were missing for each of the three variables), and these missing observations were removed from the data set. Entire cohorts with any missing data for TMORT were excluded from the analysis. After removal of these observations, 82 cohorts of pigs and 10,906 daily observations remained. Descriptive statistics for each mortality measurement were calculated and summarized using the 'base' (R Core Team, 2023) and 'dplyr' (Wickham et al., 2023) packages in R (R Core Team, 2023). Negative binomial regression models that included the fixed effect of week post-weaning and an offset of $\log(INV)$ were fit using the 'MASS' package of R (Venables and Ripley, 2002) to quantify the total found dead and euthanized pigs per number of pigs on feed for all cohorts across the feeding period, defined as the "global mortality rate curve". The R package 'ggplot2' (Wickham, 2016) was used to create all figures.

Algorithm Implementation and Performance Analysis. The algorithm was applied to the previously described dataset using R v4.3.0 (R Core Team, 2023). Sets of values from five parameters in the above algorithm were tested in a grid search experiment: 1) PADDING: {50, 100, 250, 500, 750, 1000}, 2) LAMBDA: {0, $1x10^{-10}$, $1x10^{-8}$, $1x10^{-6}$}, 3) MAGNITUDE: {0.50, 1.00, 1.50, 2.00, 2.50, 3.00}, 4) DURATION: {14, 21, 28}, and 5) PROPORTION: {0.10, 0.25, 0.50, 0.75, 0.90}. This grid expansion resulted in 2,160 combinations of the five parameters that were tested in the mortality episode classification algorithm. Three statistics for each combination of parameters were calculated to evaluate algorithm performance: 1) average found dead and euthanized pigs per day during classified mortality episodes, 2) average duration of classified mortality episodes (in days), and 3) average number of classified mortality episodes per cohort of pigs. The impact of changes in the five parameter values on the above statistics were evaluated in the context of commercial wean-to-finish pig production.

RESULTS AND DISCUSSION

Overall Mortality Curve

Figure 3.2 presents least-squares means from negative binomial regression models for the effect of week post-weaning on the total number of found dead and euthanized pigs per day per 1000 pigs in all 98 cohorts of pigs (i.e., global) and a single selected example cohort. Three major peaks were identified in the global curve at week 3, 11, and 21 post-weaning (Figure 3.2). Mehling et al. identified three different mortality patterns in an analysis of 60 lots of wean-to-finish pigs and are as follows: 1) peak mortality observed before week 17 of the feeding period, 2) low mortality (no peak) consistently observed across the feeding period, and 3) peak mortality observed after week 17 of the

feeding period, which was fewer peaks than the current study (Mehling et al., 2019). However, a single cohort of pigs will express variation around the global mortality curve in a population of pigs, in general (Figure 3.2). Parameters in the mortality episode classification algorithm (Steps 2, 3, and 8) were optimized to identify sequences of days, (minimum of 5 consecutive days and maximum of 14, 21 or 28 consecutive days) in single cohorts of pigs where mortality rate was acutely high relative to the global mortality curve (e.g., week 9 and weeks 18 to 21 of the selected cohort in Figure 3.2), which represent potential mortality episodes in commercial wean-to-finish pig production.

Parameter Grid Search Evaluation

Effect of Padding by Lambda Value. Figure 3.3 presents results for the effect of padding parameter modification by lambda parameter value on mortality episode duration, the number of mortality episodes per cohort, and the number of found dead and euthanized pigs per mortality episode day. As the padding parameter was increased from 50 to 1000 leading and lagging zero values, the average mortality episode duration linearly increased and mortality episodes per cohort linearly decreased (Figure 3.3). In general, the padding parameter had little effect on the average total found dead and euthanized pigs per mortality episode day (Figure 3.3). Padding parameter values had a direct effect on equivalent degrees of freedom per 100 observations from the penalized smoothing spline equations (**EDF**; Supplementary Figure. 3.1). As previously stated, the degrees of freedom in smoothing spline equations are a direct indicator of the flexibility of the fitted curve. For example, as equivalent degrees of freedom increase, the level of smoothing decreases. This becomes evident when evaluating the average number and

duration of classified mortality episodes per cohort. For a padding parameter value of 50, there were an average of 3.73 ± 2.786 classified mortality episodes that lasted an average of 5.05 ± 1.306 days. However, setting the padding parameter value at 1000 resulted in an average of 0.44 ± 0.531 classified mortality episodes per cohort that lasted an average of 16.31 ± 4.111 days, which is the result of less flexibility in the smoothing spline equation compared to lower padding parameter values. There were no observed differences between lambda values 0 and $1x10^{-10}$ for any statistic (Figure 3.3). However, setting the lambda parameter to $1x10^{-6}$ increased the rate of incline in average mortality episode duration across padding parameter values (Figure 3.3). In addition, at a lambda value of $1x10^{-6}$, the average number of mortality episodes per cohort and total found dead and euthanized pigs per mortality episode day were reduced compared to the smaller lambda values (Figure 3.3). The amount of penalization of smoothing spline coefficients was only impacted at lambda values greater than $1x10^{-10}$, as the equivalent degrees of freedom per 100 observations were less for lambda parameter values $1x10^{-8}$ and $1x10^{-6}$ (19.6 and 9.3 EDF, respectively; Supplementary Figure 3.2).

Effect of Magnitude by Lambda Value. The effect of modification of the magnitude parameter by lambda parameter on mortality episode duration, the number of mortality episodes per cohort, and the number of found dead and euthanized pigs per mortality episode day is given in Supplementary Figure 3.3. As the magnitude parameter increased, the average number of found dead and euthanized pigs per episode day increased, as expected (Supplementary Figure 3.3). In addition, not surprisingly, modification of the magnitude parameter had no effect on the average duration of classified mortality episodes (Supplementary Figure 3.3). However, changes in the

magnitude parameter had the greatest effect on the average number of classified mortality episodes per cohort. Across all lambda parameter values, as the magnitude parameter increased, the number of classified mortality episodes per cohort decreased (Supplementary Figure 3.3). Across magnitude parameter values, there were no observed differences between lambda parameter values for the average found dead and euthanized pigs per mortality episode day (Supplementary Figure 3.3). Similar values were observed for all three statistics at 0, $1x10^{-10}$, and $1x10^{-8}$ (Supplementary Figure 3.3). At a lambda parameter of $1x10^{-6}$, fewer mortality episodes per cohort were classified, and these mortality episodes lasted longer than classified mortality episodes at the other lambda parameter values (Supplementary Figure 3.3). The magnitude parameter simply provides a threshold to declassify "first pass" mortality episodes (see section "Algorithm to Detect Mortality Episodes") that fail to meet the selected average number of found dead and euthanized pigs per day during the episode. Therefore, it has no effect on the smoothing spline regression equation, compared to the padding and lambda parameters. Higher selected magnitude parameter values will provide a stricter classification of mortality episodes (i.e., less classified episodes per cohort). For example, setting the magnitude parameter at 0.50 and 3.00 total dead pigs per mortality episode day produced an average of 3.59 ± 3.354 found dead and euthanized pigs per day and 0.66 ± 0.535 mortality episode classifications per cohort across all grid search experiment combinations, respectively. Thus, wean-to-finish production systems that suffer from high mortality may benefit from higher magnitude parameter values, while production systems that observe a lower mortality rate relative to industry standards may benefit from selecting lower magnitude parameter values.

Effect of Duration by Lambda Value. Supplementary Figure 3.4 presents results for the effect of duration parameter modification by lambda parameter value on mortality episode duration, the number of mortality episodes per cohort, and the number of found dead and euthanized pigs per mortality episode day. Of all the parameters in the grid search experiment, the duration parameter had the least effect on each mortality episode statistic, as expected (Supplementary Figure 3.4). There were no observed differences across the duration parameter values for the average number of found dead and euthanized pigs and mortality episodes per cohort (Supplementary Figure 3.4). However, as the duration parameter value increased the average mortality episode duration increased, but these changes were relatively small (Supplementary Figure 3.4). Like the other parameters, setting the lambda parameter value to 1x10-6 generally caused the algorithm to classify longer but fewer mortality episodes per cohort, on average, compared to the smaller lambda parameter values (Supplementary Figure 3.4). The purpose of the duration parameter in the mortality episode classification algorithm is to allow users to constrain the length of classified mortality episodes. Mortality episodes that are extremely short (i.e., less than 5 days) tend to be influenced by singular days with abnormally high mortality. On the other hand, classified mortality episodes longer than 28 days tend to represent chronic as opposed to acute mortality. There are standard protocols in place to mitigate the effects of chronic mortality in wean-to-finish pig barns, such as feed and water antibiotics or zone-heating during critical periods of high vulnerability. However, intense periods of acute mortality are harder to identify and diagnose, which is the goal of the proposed mortality episode classification algorithm.

Effect of Proportion by Lambda Value. Supplementary Figure 3.5 shows results for the effect of proportion parameter modification by lambda parameter value on mortality episode duration, the number of mortality episodes per cohort, and the number of found dead and euthanized pigs per mortality episode day. As expected, there were no differences in mortality episode duration across proportion parameter values (Supplementary Figure 3.5). The number of found dead and euthanized pigs per mortality episode day was similar for proportion parameter values 0.10 to 0.75 but was higher for proportion parameter value 0.90 (Supplementary Figure 3.5). Modification of the proportion parameter had the largest effect on the number of mortality episodes per cohort. For example, as the proportion parameter value increased, the number of mortality episodes per cohort decreased, especially after values greater than 0.25, which could be considered a breakpoint in this relationship (Supplementary Figure 3.5). Like the other parameters described previously, the lambda parameter value only influenced mortality episode duration and the number of mortality episodes per cohort at a selected value of $1x10^{-6}$ (Supplementary Figure 3.5). Inclusion of the proportion parameter in the mortality episode classification algorithm ensures that episodes with one or two singular "outlier" days are not included in the final classification. For example, if a classified mortality episode has a duration of 10 days at a proportion parameter value of 0.50, then at least 1 dead pig was observed on at least 5 of the days within the episode. However, a mortality episode could be classified that has a duration of 10 days where 10 dead pigs were observed on a single day without the use of the proportion parameter (assuming a magnitude parameter value of 1). As previously defined, mortality episodes in this analysis are sequences of days where higher than average mortality relative to the cohort

baseline at a given pig age is observed and sustained for 5 to 14, 21, or 28 days (depending on duration parameter value). Thus, the latter example would fail to adequately meet this definition.

Selection of Optimal Parameter Values

In the selection of optimal parameter values for the mortality episode classification algorithm, careful evaluation and manual inspection of resulting classifications is initially required. As demonstrated above, small changes to any one parameter could cause variation in algorithm behavior. In addition, there will be differences across wean-to-finish production systems or companies in mortality episode classification due to varying levels of mortality and diverse wean-to-finish standard operating procedures. Nonetheless, the selected parameter values should satisfy the following criteria:

1) Classified mortality episodes represent acute as opposed to chronic mortality.
2) Mortality is sustained for the duration of the classified mortality episode.
3) Average mortality per day during classified mortality episodes is significantly higher than average mortality per day in days excluded from classified mortality episodes.

In this study, we used results from grid search experiments across combinations of parameter values to inform the selection of optimal parameters. Figure 3.4 presents an example of resulting mortality episode classification using optimized parameters in a selected cohort of wean-to-finish pigs. In this example, the following parameter values were selected: 1) PADDING: 625, 2) LAMBDA: 1x10-8, 3) MAGNITUDE: 1, 4) DURATION: 21, and 5) PROPORTION: 0.50. The resulting smoothed curve has

adequate flexibility to properly model potential mortality episode peaks across the wean-to-finish period (Figure 3.4). In addition, the labeled "start" to each mortality episode (green dashed line; Fig. 4) precisely identifies the day in which mortality begins to increase (Figure 3.4). Moreover, the algorithm does not classify a mortality episode at approximately day 30 post-placement, where 8 dead pigs were observed on a single day with limited mortality on surrounding days, which represents an outlier in this cohort (Figure 3.4). Modifying the padding parameter drastically impacted the resulting mortality episode classifications in this cohort (Supplementary Figure 3.6). Selection of a higher padding parameter value (1000) resulted in too few episode classifications with longer durations (Supplementary Figure 3.6A), while selection of a lower parameter value (250) resulted in far too many mortality episode classifications with short durations (Supplementary Figure 3.6C). Furthermore, similar effects on the resulting smoothed curve and classified mortality episodes were observed upon modification of the lambda parameter (Supplementary Figure 3.7). For example, increasing the smoothing penalization with a lambda parameter value of $1x10^{-6}$ resulted in a rigid smoothed curve and zero classified mortality episodes (Supplementary Figure 3.7A). However, decreasing the lambda parameter value to 0 produced a smoothed curve (Supplementary Figure 3.7C) that was nearly the same as the optimal smoothed curve determined by visual inspection (Supplementary Figure 3.7B), which was consistent with the results from the grid search experiment (see section "Parameter Grid Search Evaluation"). Upon manual visualization of similar figures for each cohort, we recommend the use of parameter values from Figure 3.4 as starting values when applying this algorithm in

future wean-to-finish cohorts. However, it must be stated that modification of each parameter may be required to best suit individual production systems.

Implications in the Commercial Swine Industry

There are multiple potential use cases for refined and automated mortality episode classifications in the commercial swine industry. The most applicable to modern production is utilization of this algorithm to label the start of mortality episodes for further use as a dependent variable in advanced statistical and machine learning prediction models. Precise prediction of the start of a mortality episode provides several advantages to wean-to-finish pig producers. First, with adequate delay between prediction and the actual start of a predicted mortality episode, targeted intervention plans consisting of broad-use antibiotics or increased individual pig care can be formulated by veterinarians or pig managers. Predictive models for real-time daily mortality require large datasets that span hundreds or thousands of individual cohorts. Through this algorithm, we provide a method to reproducibility and efficiently label training data to be used in machine learning models. In addition, we are unaware of a published method of programmatically labeling mortality episodes using a well-defined set of rules that are based in statistical evaluation. Furthermore, this algorithm can be used in a retrospective fashion in the production and research setting. For example, “barn closeout” records are common practice in commercial production. The addition of episodic patterns of mortality to historical closeout records provides additional information on the efficacy of health management practices in previous cohorts. In the research setting, the effect of certain antibiotic, vaccine, or nutritional protocols on mortality episode patterns can be

evaluated on a more refined level. Finally, this approach could be easily modified to other livestock production systems, such as broiler farms or beef feedlots.

CONCLUSIONS

In this paper, we present a novel algorithm to classify mortality in commercial wean-to-finish pig barns. The algorithm, which is based on penalized smoothing spline regression equations, precisely identifies sequences of days with sustained increased mortality relative to the baseline of individual wean-to-finish pig cohorts. Adoption of this algorithm in the swine industry could positively impact mortality levels through use as a labeling mechanism for machine learning training data or as a performance evaluation metric in general pig production and research.

FIGURES

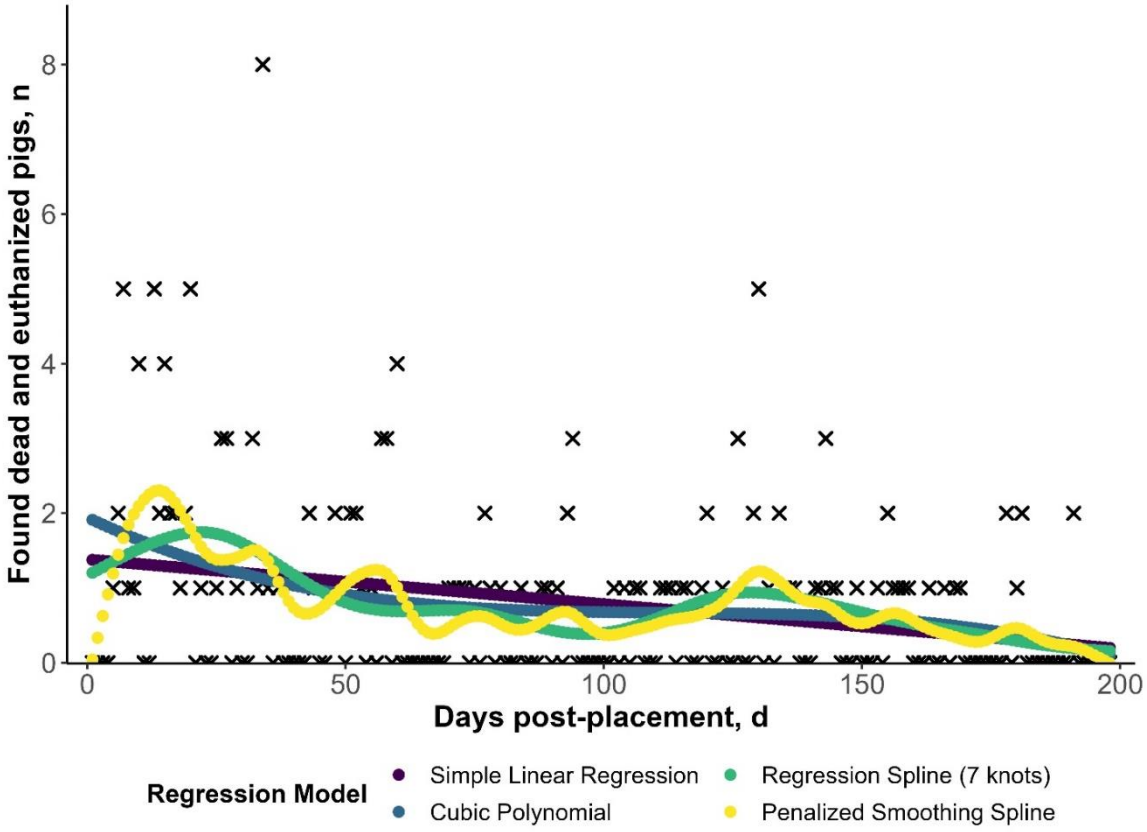


Figure 3.1. Comparison of four regression methods for smoothing the relationship between days post placement and number of daily found dead and euthanized pigs in a selected cohort.

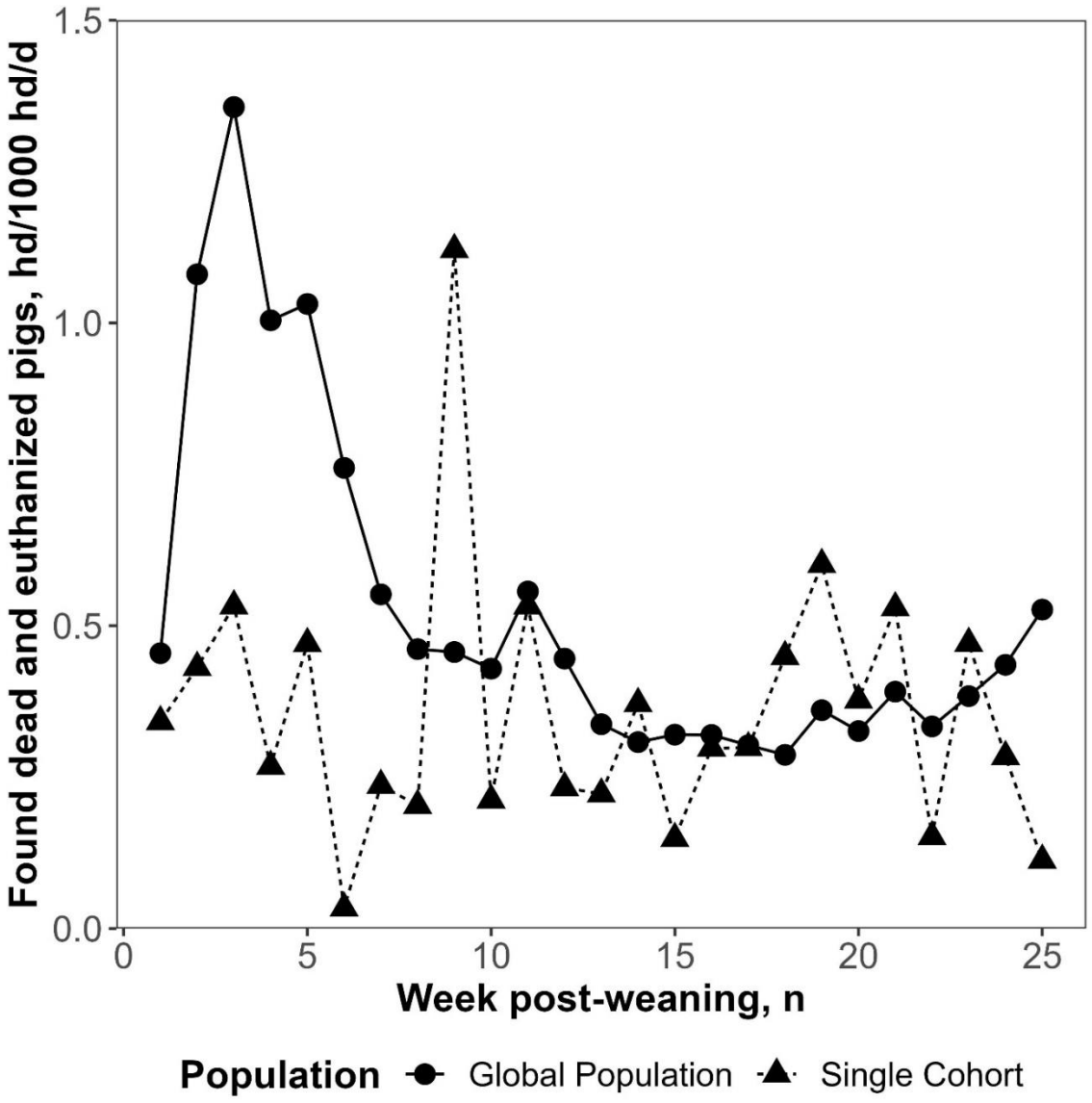


Figure 3.2. Mortality rate curves between week post-weaning and number of found dead and euthanized pigs per 1000 pigs for the entire data set (Global Population) and a selected cohort (Single Cohort).

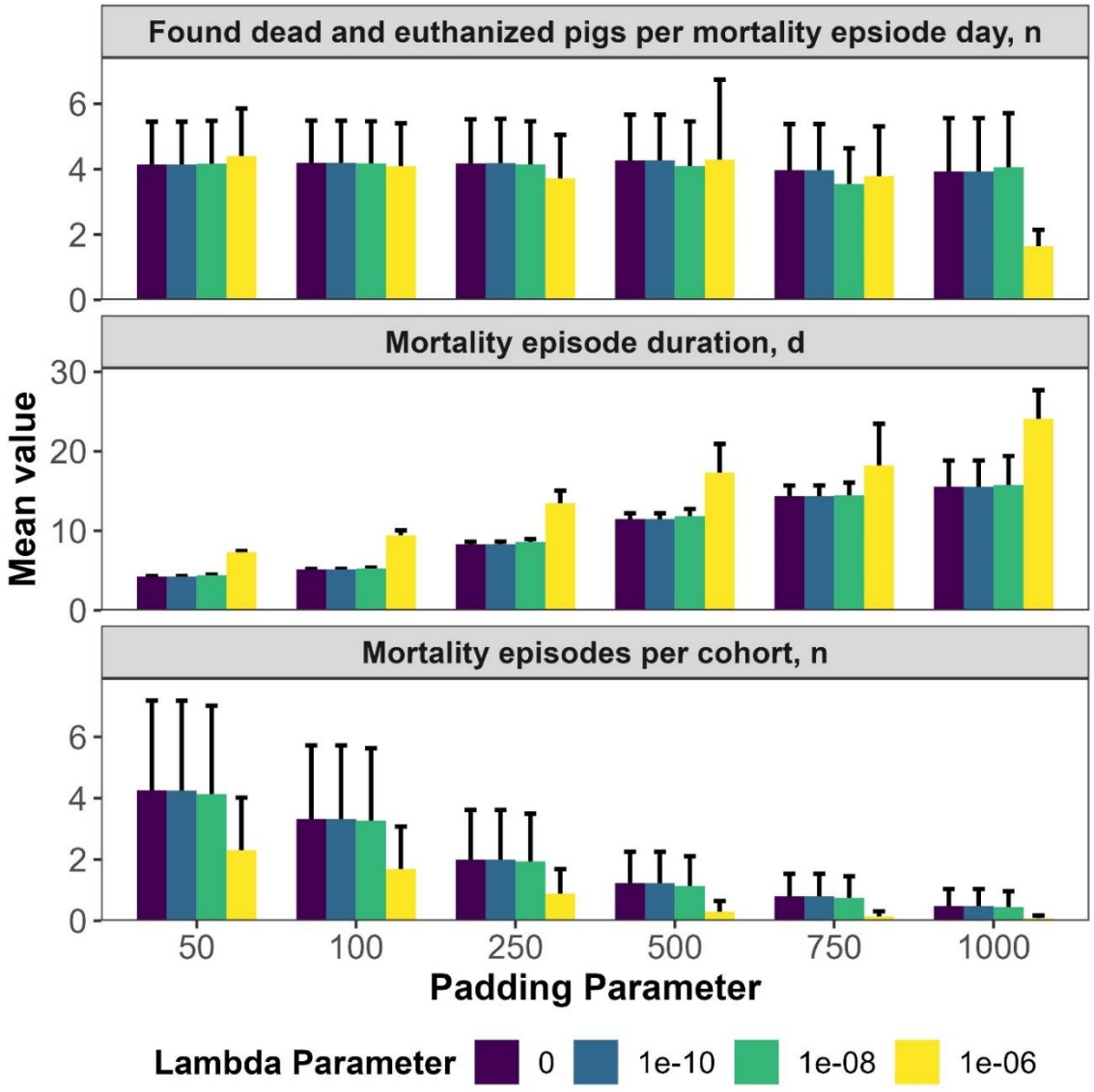


Figure 3.3. Results from parameter grid search for the effect of padding and lambda parameters on the number of found dead and euthanized pigs per mortality episode day, mortality episode duration, and the number of mortality episodes per cohort. Bars represent the mean value for each statistic across all experiments using a specific padding and lambda value. Standard error bars represent +/- one standard deviation for all values for each statistic at a specific padding and lambda value.

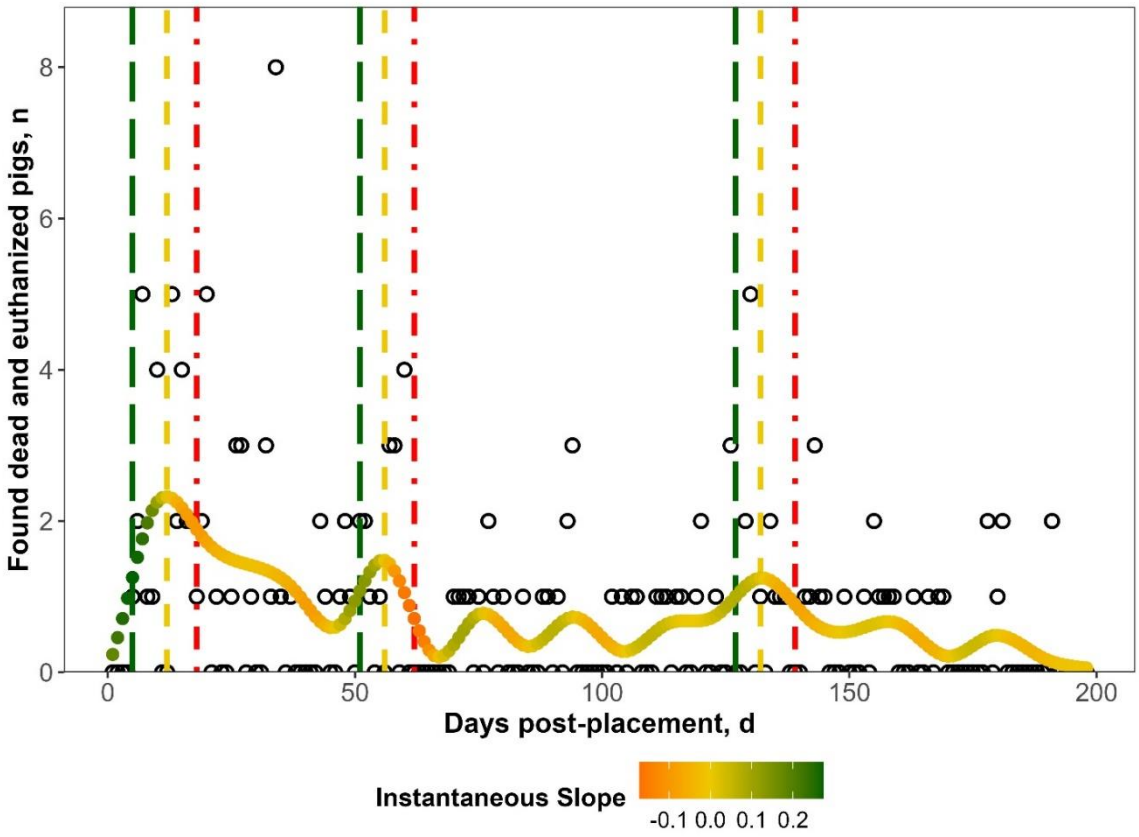
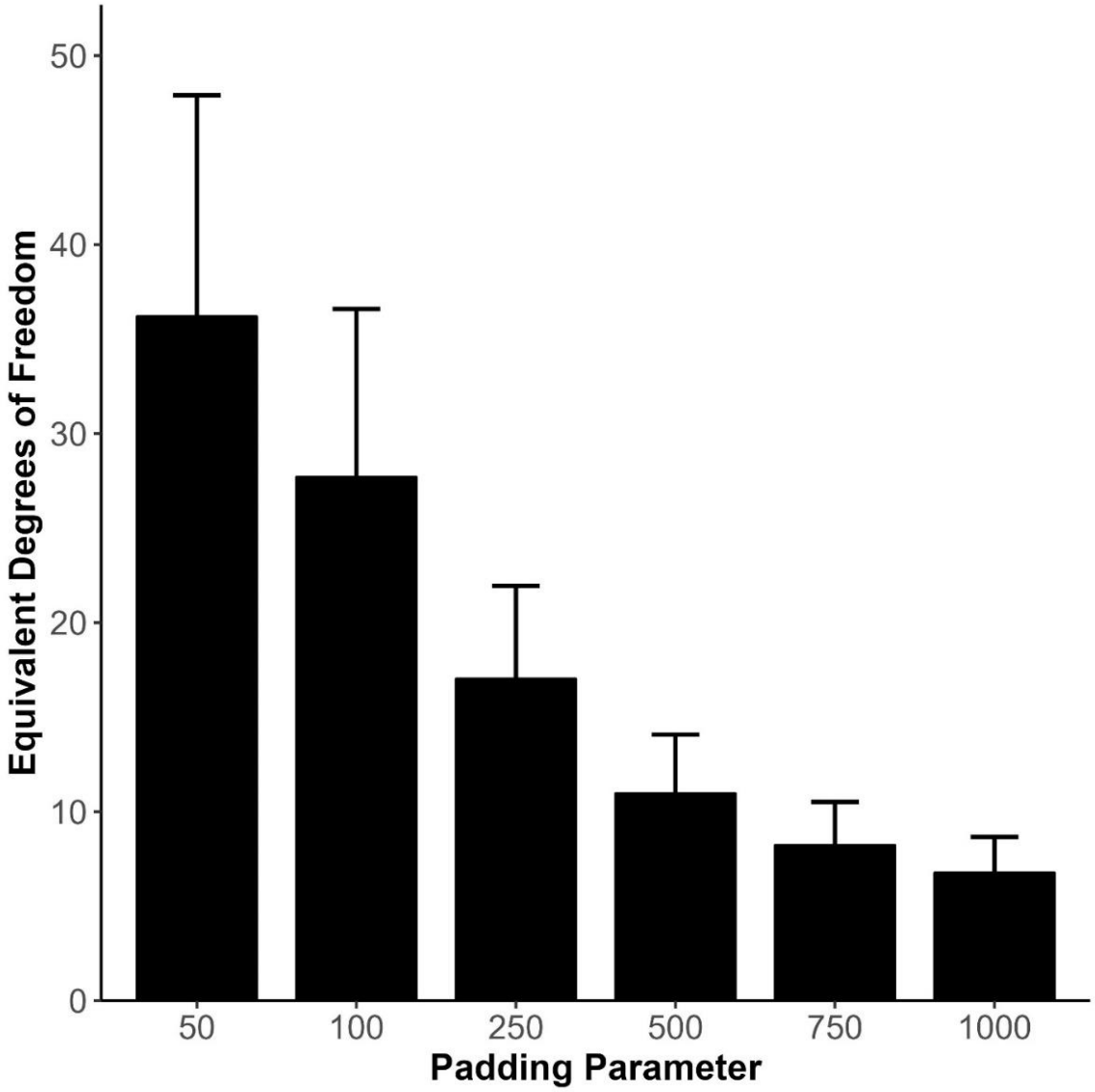
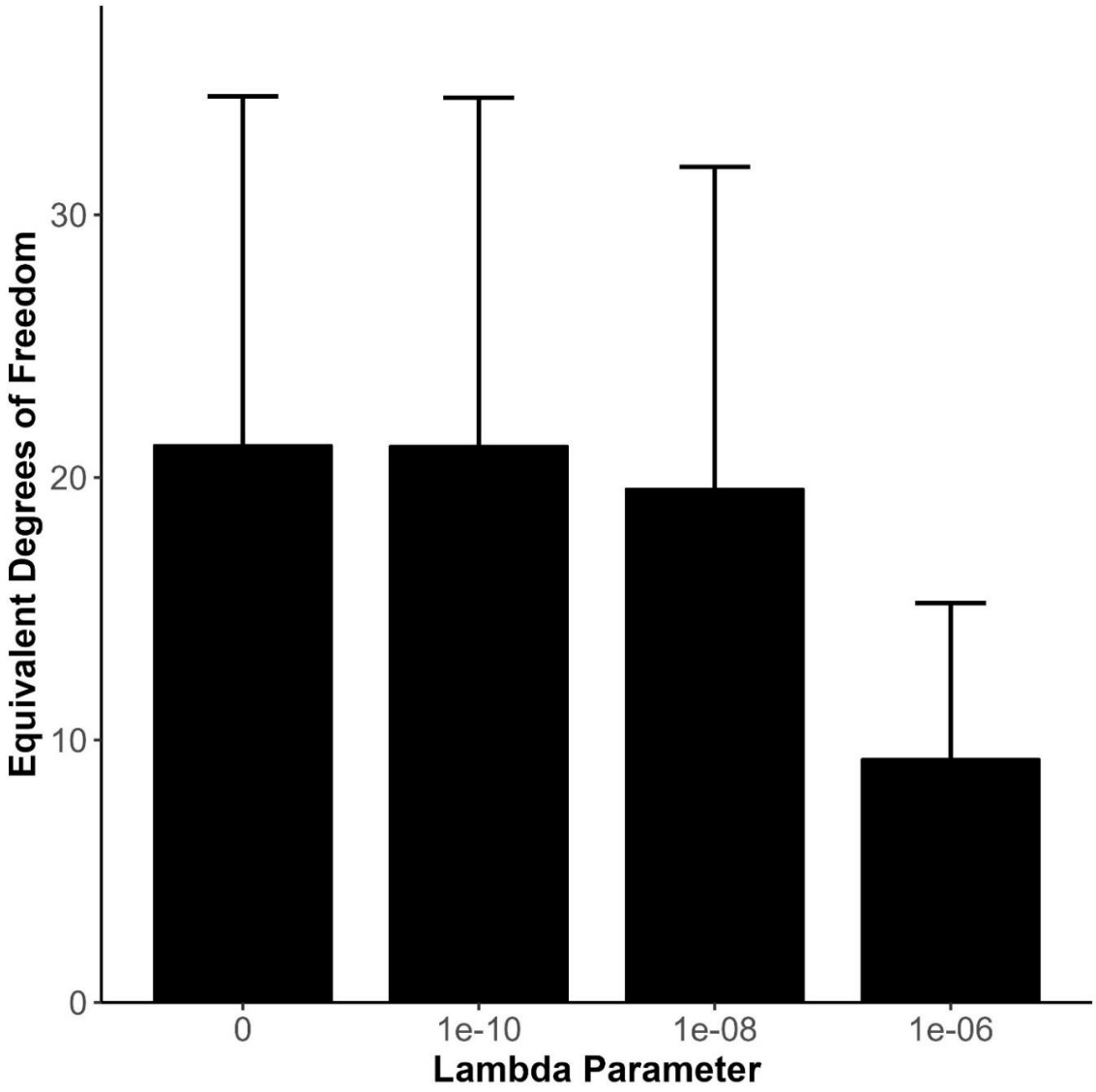


Figure 3.4. Mortality episode classifications for a selected cohort for optimal selected parameter values (padding = 625, lambda = $1x10^{-8}$, magnitude = 1, duration = 21, and proportion = 0.50) in the mortality classification algorithm. Vertical dashed green, yellow, and red bars represent the start, peak, and end of each classified mortality episode, respectively. Empty black circles are the observed number of found dead and euthanized pigs per day, while the solid-colored circles are the predicted number of found dead and euthanized pigs per day from the penalized smoothing spline equation.

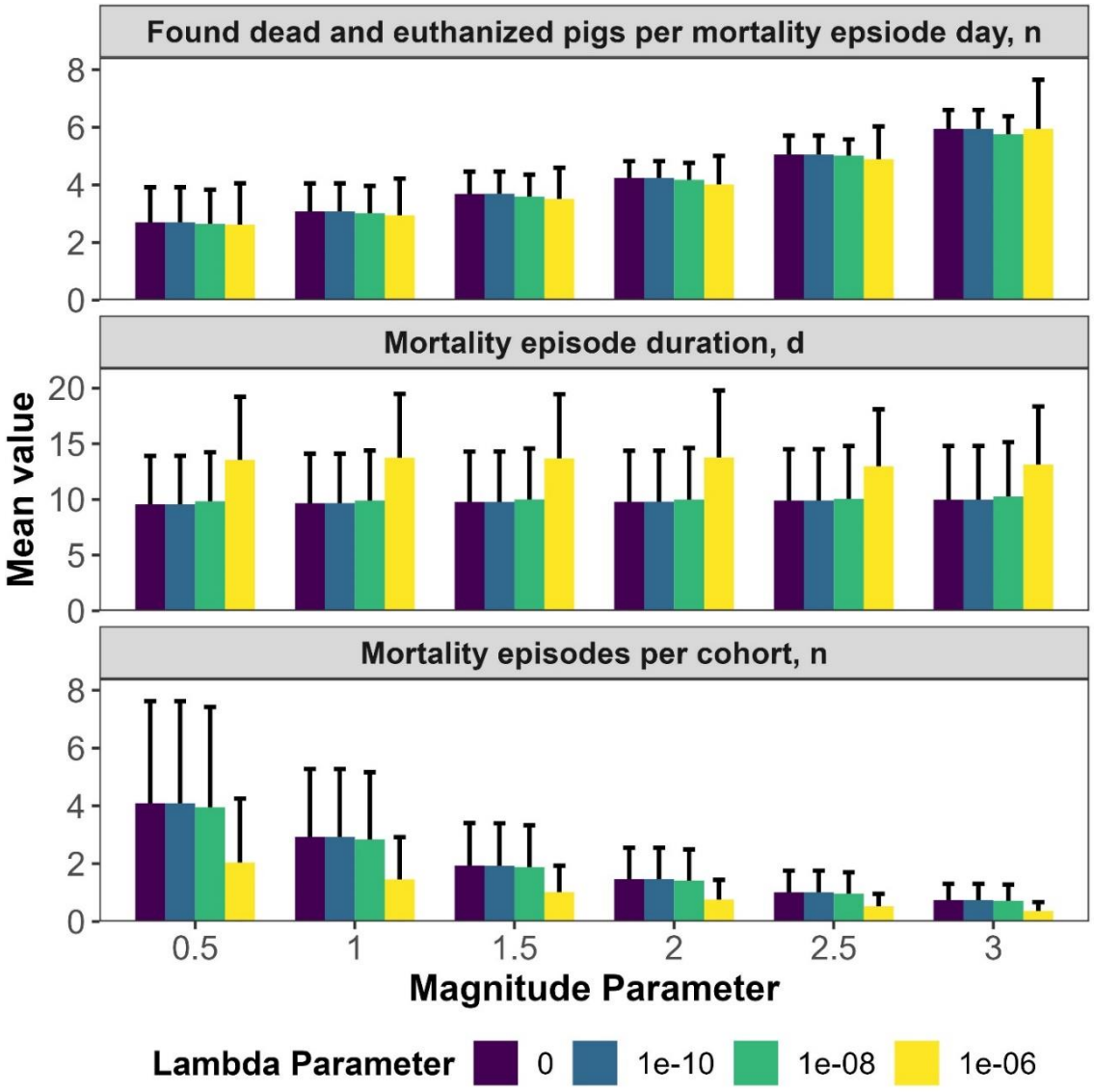
SUPPLEMENTARY FIGURES



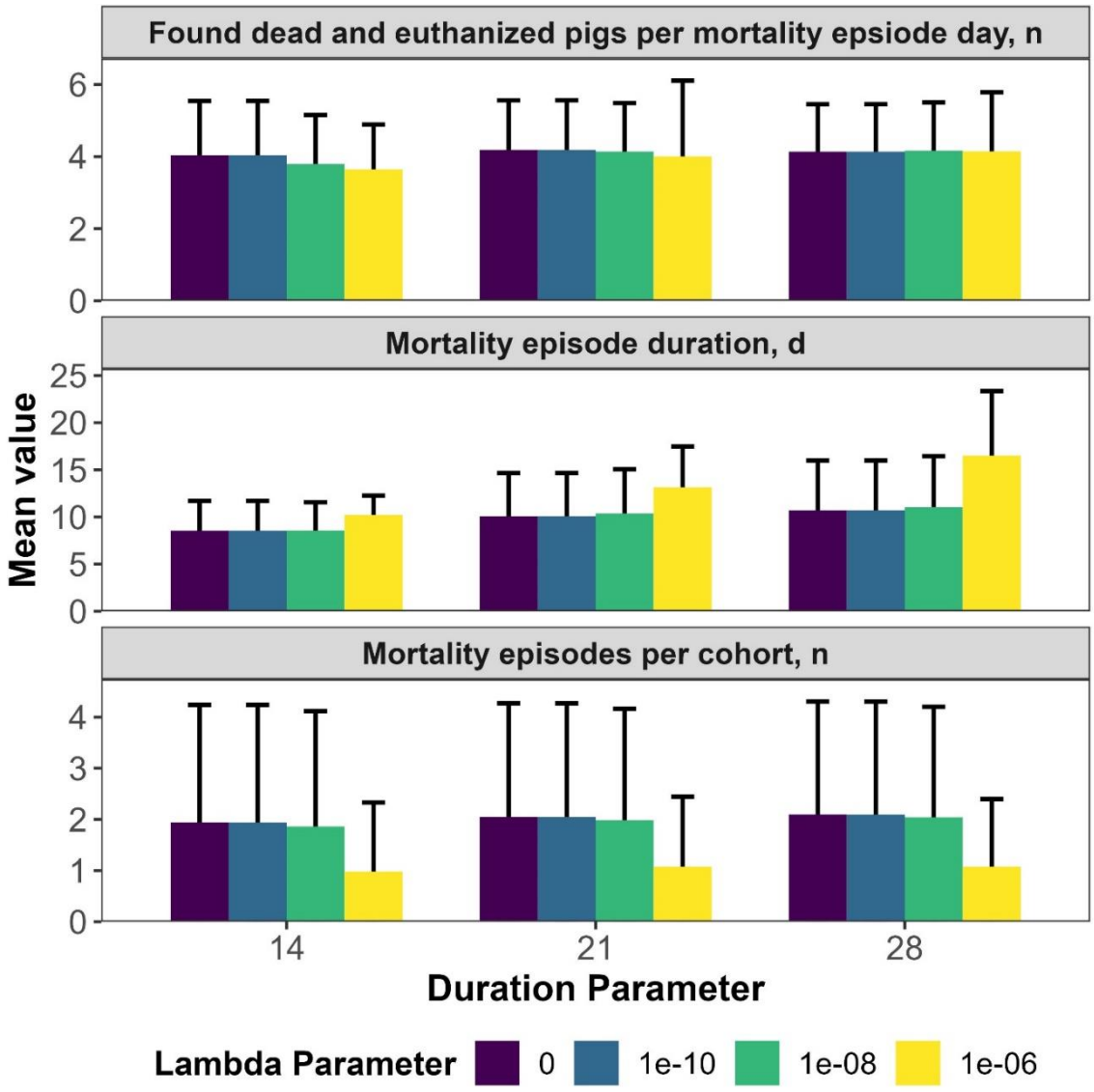
Supplementary Figure 3.1. Relationship between equivalent degrees of freedom from penalized smoothing spline regression equations and padding parameter values. Bars represent the mean value across all experiments at each padding value. Error bars represent +/- one standard deviation for all mean values of equivalent degrees of freedom estimated for each parameter value.



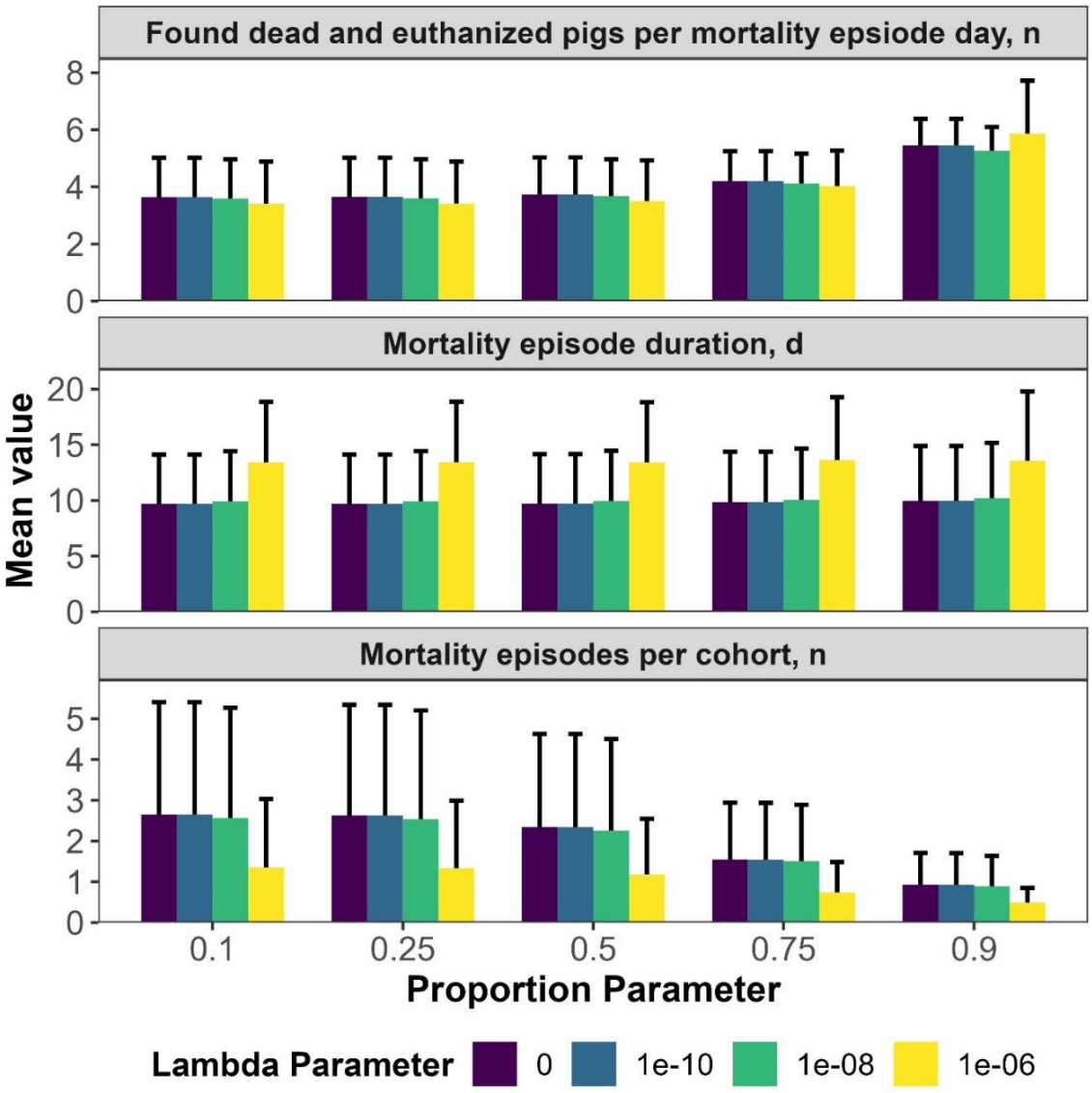
Supplementary Figure 3.2. Relationship between equivalent degrees of freedom from penalized smoothing spline regression equations and lambda parameter values. Bars represent the mean value across all experiments at each lambda value. Error bars represent +/- one standard deviation for all mean values of equivalent degrees of freedom estimated for each lambda value.



Supplementary Figure 3.3. Results from parameter grid search for the effect of magnitude and lambda parameters on the number of found dead and euthanized pigs per mortality episode day, mortality episode duration, and the number of mortality episodes per cohort. Bars represent the mean value for each statistic across all experiments using a specific magnitude and lambda value. Error bars represent +/- one standard deviation for all values for each statistic at a specific magnitude and lambda value.

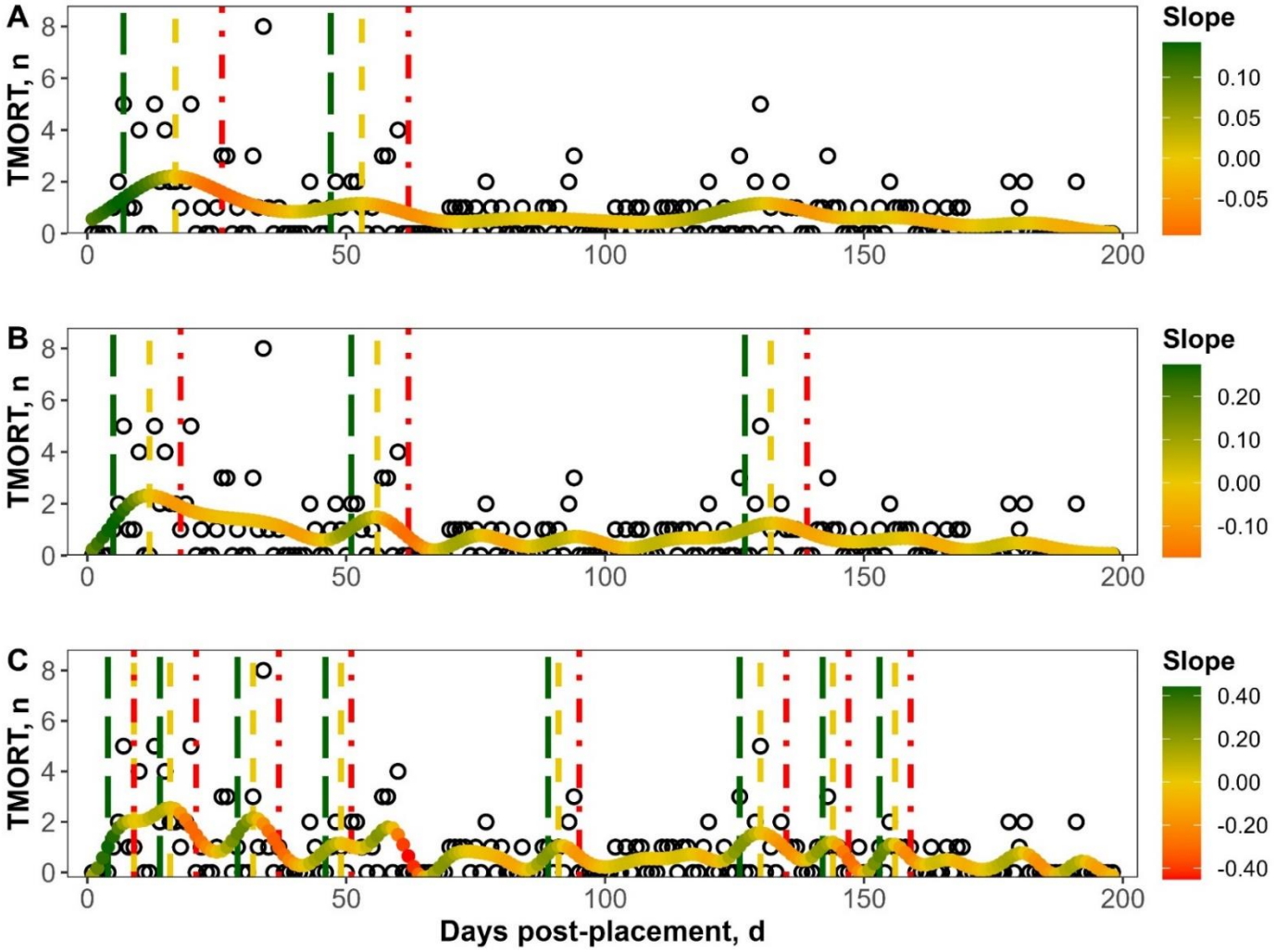


Supplementary Figure 3.4. Results from parameter grid search for the effect of duration and lambda parameters on the number of found dead and euthanized pigs per mortality episode day, mortality episode duration, and the number of mortality episodes per cohort. Bars represent the mean value for each statistic across all experiments using a specific duration and lambda value. Error bars represent +/- one standard deviation for all values for each statistic at a specific duration and lambda value.



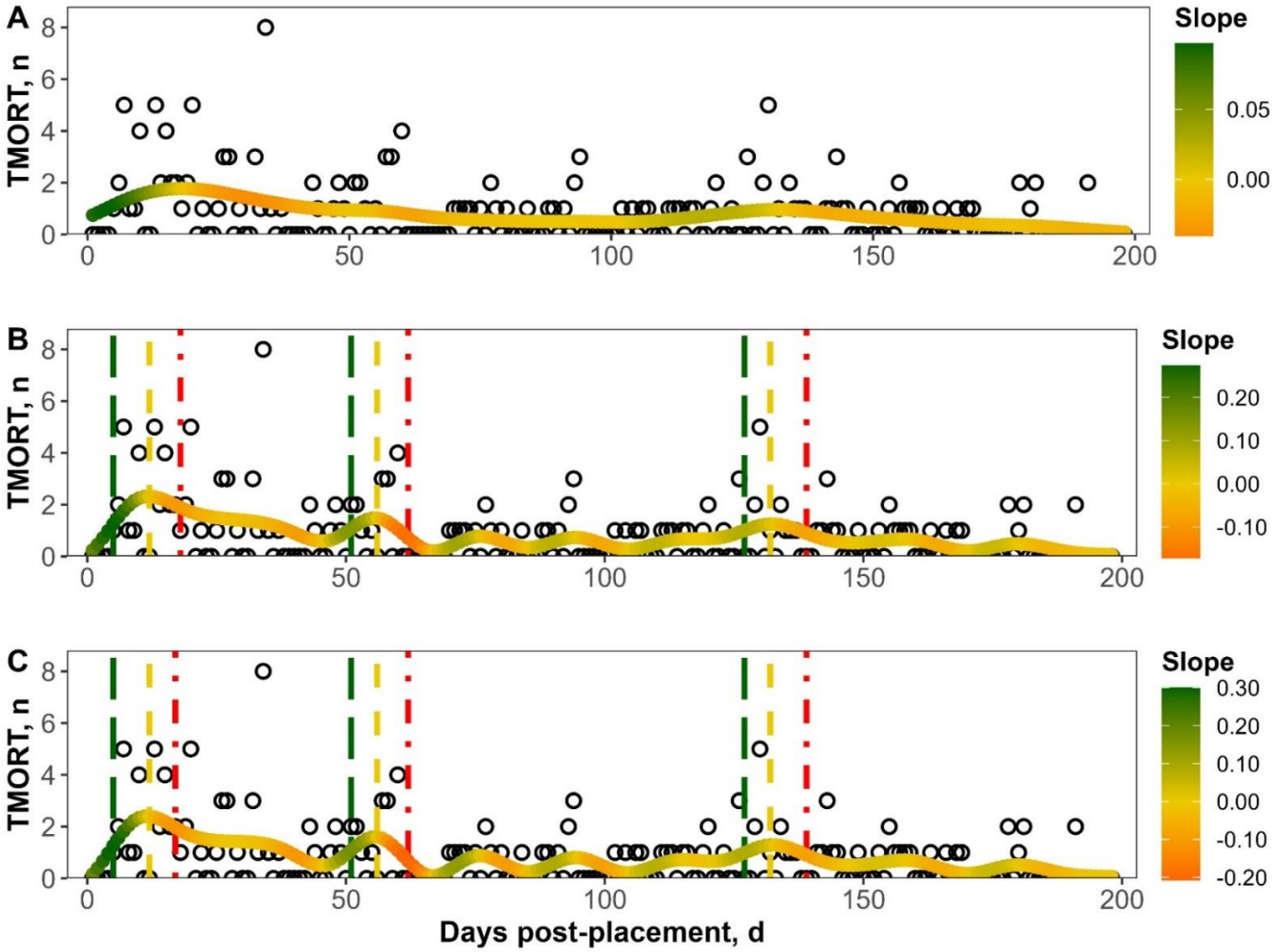
Supplementary Figure 3.5. Results from parameter grid search for the effect of proportion and lambda parameters on the number of found dead and euthanized pigs per mortality episode day, mortality episode duration, and the number of mortality episodes per cohort. Bars represent the mean value for each statistic across all experiments using a specific proportion and lambda value. Error bars represent +/- one standard deviation for all values for each statistic at a specific proportion and lambda value.

Supplementary Figure 3.6



Supplementary Figure 3.6. Mortality episode classifications for a selected cohort for (A) padding parameter =1000 and other parameters equal to optimal parameters, (B) optimal selected parameter values (padding = 625, lambda = $1x10^{-8}$, magnitude = 1, duration = 21, and proportion = 0.50), and (C) padding parameter = 250 and other parameters equal to optimal parameters in the mortality classification algorithm. Vertical dashed green, yellow, and red bars represent the start, peak, and end of each classified mortality episode, respectively. Empty black circles are the observed number of found dead and euthanized pigs per day, while the solid-colored circles are the predicted number of found dead and euthanized pigs per day from the penalized smoothing spline equation.

Supplementary Figure 3.7



Supplementary Figure 3.7. Mortality episode classifications for a selected cohort for (A) lambda parameter =$1x10^{-6}$ and other parameters equal to optimal parameters, (B) optimal selected parameter values (padding = 625, lambda = $1x10^{-8}$, magnitude = 1, duration = 21, and proportion = 0.50), and (C) lambda parameter = 0 and other parameters equal to optimal parameters in the mortality classification algorithm. Vertical dashed green, yellow, and red bars represent the start, peak, and end of each classified mortality episode, respectively. Empty black circles are the observed number of found dead and euthanized pigs per day, while the solid-colored circles are the predicted number of found dead and euthanized pigs per day from the penalized smoothing spline equation.

CHAPTER 4. MACHINE LEARNING MODELS TO PREDICT REAL-TIME MORTALITY OUTCOMES IN COMMERCIAL WEAN-TO-FINISH PIG BARNS

Caleb J. Grohmann[†], Caleb M. Shull[ψ], Erin J. Lowe[¥], Dale D. Polson[ϕ€], and Jared E. Decker [†‡§*]

[†]Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA

[‡]Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA

[§]Genetics Area Program, University of Missouri, Columbia, MO 65211, USA

[ψ]The Maschhoff's, LLC, Carlyle, IL 62231, USA

[¥]Summit Smart Farms, Remington, IN 47977, USA

[ϕ]Boehringer Ingelheim Vetmedica GmbH, Ingelheim am Rhein, Germany

[€]SoundTalks NV, Leuven, Belgium

*Corresponding author

LAY SUMMARY

Wean-to-finish mortality in pigs has been increasing over the past several years. Data from a diverse set of sensors can be used to forecast mortality outcomes during the growing phase of pigs. In this study, models were tested for the ability to forecast mortality events in commercial wean-to-finish barns. The highest performing model proved to be precise and conservative by reporting precision scores that were higher than a baseline model and false positive rates that were low. Data from water disappearance and cough incidence sensors were most important for accurate predictions. Models in this study can be used to promote data-driven decision making by alerting farmers to periods of time where preventative interventions are required.

TEASER TEXT

Machine learning models to predict episodic and sporadic mortality outcomes in commercial wean-to-finish pig barns were evaluated. In this study, the most complex model achieved high precision relative to a baseline classifier and other simpler models and low false positive rates when evaluated on unseen data. Models in this study can be used to promote data-driven decision making by alerting farmers to periods of time where preventative interventions are required.

ABSTRACT

Precision livestock farming is increasing in popularity year after year in the swine industry due to the availability of innovative sensors, microphones, and cameras for automatic surveillance of pig barns. Wean-to-market mortality in pigs is of utmost economic importance to pig farmers, especially as mortality rates have increased, in general. Machine learning models provide an interesting opportunity to combine data from multiple sources into a single model to predict real-time daily mortality outcomes in growing pigs. In this study, we evaluated the ability of three machine learning frameworks [elastic net logistic regression, support vector machine, and gradient boosted decision trees (i.e., XGBoost)] to forecast sporadic and episodic mortality outcomes. These models were trained and tested on 5,364 daily observations from 6 commercial wean-to-finish farms owned and operated by The Maschhoff's, LLC (Carlyle, Illinois, USA). Features used in the analysis were from manually recorded production data, SoundTalks® cough sensors (SoundTalks NV, Leuven, Belgium), and ventilation controller systems that collected water disappearance and climatic conditions. Across all models, prediction of episodic mortality was more feasible than sporadic mortality [Cross-validation F0.5 scores = 0.30 to 0.51 for episodic (**MEP**) and 0.16 to 0.24 for sporadic (**EHMD**) mortality]. Out of all models, the most complex model, XGBoost, reported the highest F0.5 scores during cross-validation and holdout testing for mortality episode prediction (0.51 and 0.40, respectively), which was considerably better than the naïve baseline classifier. In addition, XGBoost proved to be a conservative classifier for mortality episodes, reporting false positive rates of 0.08 and 0.14 during cross-validation and holdout testing, respectively. In this model, age, water disappearance, and

SoundTalks® respiratory health status achieved the highest variable importance for episodic mortality (0.26, 0.17, and 0.08 relative accuracy gain, respectively). Receiver operating and precision-recall curves were used to evaluate the overall classification ability of each model based on unseen data from subsequent pig cohorts. Each model achieved a high area under the receiver operating curve (AUC = 0.77 to 0.78), which were considered accurate classifiers. Area under the precision-recall curve for each model was moderately high compared to the naïve baseline classifier (AUCPR = 0.30 to 0.46; baseline classifier = 0.25). Models from this analysis can promote data-driven decision making regarding preventative interventions to reduce mortality losses in commercial wean-to-finish pig barns.

INTRODUCTION

In the age of artificial intelligence and the Internet of Things (IoT), farmers have access to unprecedented amounts of data originating from a diverse set of sources. Sensors, microphones, cameras have enabled automatic and real-time surveillance of farms, which aid farmers in the management of livestock, often termed precision livestock farming (Wathes et al., 2008; Berckmans and Guarino, 2017; Vranken and Berckmans, 2017; Morota et al., 2018). These technologies provide a refined view of animal health, behavior, and nutrient consumption and/or growth alongside environmental conditions such as temperature and air quality (Piñeiro et al., 2019). Thus, on these farms, evidence-based decision-making can be utilized to improve feed efficiency, morbidity and mortality, and other economically relevant outcomes at the farm or animal level (Morota et al., 2018; Van Klompenburg and Kassahun, 2022).

Intense monitoring of animal health and well-being, enabled by sensor technology, has grown in relevance in pig production over the past several years (Matthews et al., 2017). In this area, wean-to-finish mortality is of critical importance due to recent increases observed from 2017 to 2021 in the pork industry and associated impacts on producer profitability (MetaFarms and National Pork Board, 2022) and pig welfare. Mortality outcomes in growing pigs are influenced by a complex array of causes such as infectious diseases and environmental, animal, and management factors, along with numerous potential interactions (Gebhardt et al., 2020a; Gebhardt et al., 2020b). Barns outfitted with various state-of-the-art sensors can track mortality indicators at unprecedented levels, which allows farmers and researchers the chance to disentangle this

web of causal factors. However, advanced analytical approaches are required to maximize success in this area of precision swine farming.

Machine learning models provide an interesting opportunity to further understand mortality outcomes. Because these models can integrate large amounts of data from countless mortality indicators, learn from past patterns, and automatically estimate higher order interactions and nonlinearities, there is potential for use as a method to forecast real-time mortality in growing pigs. Accurate and timely predictions of mortality enable preventative interventions and a reduction of animal losses (Lasser et al., 2021). Unfortunately, very little research exists on the evaluation of machine-learning models to predict mortality in pigs. Therefore, the objective of the current study was to evaluate the ability of three different machine learning model frameworks [elastic net logistic regression (**ELNR**), support vector machines (**SVM**), and gradient boosted decision trees (i.e., XGBoost; **XGB**)] in the real-time prediction of episodic and sporadic weaning-to-finishing pig mortality using data derived from pig health and environmental sensor technology.

MATERIALS AND METHODS

Animal Care and Use Committee approval was not needed because the data was obtained from a preexisting commercial database.

Animals and Facilities

The data for all analyses in the current study was collected from 98 different cohorts of pigs grown at six different wean-to-finish sites owned and operated by The Maschhoff's, LLC (Carlyle, Illinois, USA) from December 2020 to October 2023 (sites A, B, C, D, E, and F). Detailed descriptions of each cohort of pigs are provided in

Supplementary Table 4.1. A cohort of pigs was considered a unique group of pigs allotted to a single room within a single site. Each of the six wean-to-finish sites contained two rooms of pigs under study, which totaled twelve rooms across the entire trial. Three of the six sites were in Iowa and were considered commercial wean-to-finish sites, where only standard pig feeding and growth occurred. Rooms in each commercial site were 50 feet × 193 feet in dimension and contained fourteen separate pens. The other three sites in the current trial were in central and southern Illinois and were considered commercial research wean-to-finish sites. In these research sites, cohorts of weaning pigs were routinely allotted and evaluated for growth performance and mortality measures until the end of the finishing phase as a part of nutrition, health, and genetics trials relevant to commercial pig production. However, these trials were expected to have minimal impact on data collected in the current study. Rooms in each of the research sites were 50 feet × 300 feet in dimension and contained 60 to 90 pens.

Pigs in each cohort were sired by an AcuFast™ (Saskatoon, Saskatchewan, Canada; formerly Acuity™, Carlyle, Illinois, USA) Duroc terminal sire bred to a Yorkshire × Landrace maternal dam, mixed sex, and weaned at twenty to twenty-one days of age. All sites in this trial were managed as “all-in, all-out”, and each room was populated within seven days to reduce health concerns between and within cohorts of pigs. During the nursery period, which was defined as week zero post-weaning until week seven to twelve post-weaning, pigs were “double-stocked” to approximately three-square feet per pig, which is standard protocol within The Maschhoff’s system. “Thin-down” occurred between week seven and twelve post-weaning for all cohorts, which is defined as the process of removing half of the pigs from each pen to be raised in a

different location for the duration of the growing-finishing period. Any pigs removed from a cohort during the thin-down process were subsequently removed from the present study. After thin-down, all pigs were stocked to a density of six square feet per pig. Finishing pigs, upon reaching a fixed market weight of approximately 285 pounds, were sent to harvest over a period of approximately 6 to 8 weeks. Animal housing, feeding, handling, and veterinary care were under the supervision of The Maschhoff's management personnel. All rooms had fully slatted floors, deep-pit manure handling, mechanically controlled ventilation, and automated feeding and bowl waterers. Pigs were provided *ad libitum* access to feed and water from weaning to harvest in a wet-dry feeding system and were fed standard commercial corn-soybean pig diets.

All pigs received standard vaccination and medications that followed The Maschhoff's standard protocol, which was detailed by Krahn as follows: *Mycoplasma hyopneumoniae* vaccine (Fostera® Gold PCV MH, Zoetis, Kalamazoo, MI, USA; Circumvent® PCV-M G2, Merck Animal Health, Summit, NJ, USA or Ingelvac MycoFlex®, Boehringer Ingelheim Vetmedica Inc, St. Joseph, MO, USA) at processing (3 to 5 days of age), and at 2-weeks post-weaning, porcine reproductive respiratory syndrome virus modified-live virus vaccine (Ingelvac PRRS® MLV, Boehringer Ingelheim Vetmedica Inc, St. Joseph, MO, USA) at 2-weeks post-weaning, and porcine circovirus type 2 (PCV2) killed vaccine (Fostera Gold PCV® MH, Zoetis, Kalamazoo, MI, USA; Circumvent® PCV-M G2, Merck Animal Health, Summit, NJ, USA or Ingelvac CircoFlex®, Boehringer Ingelheim Vetmedica Inc, St. Joseph, MO, USA) vaccine at 3-weeks post-weaning (Krahn, 2018). Feed medication protocol followed The Maschhoff's standard protocols and were kept consistent between all groups of pigs. All

water and injectable antimicrobial treatments and interventions performed were part of the routine care administered to animals by their caretakers.

Data Collection and Calculated Variables

Measurements for each room within site were collected by either barn personnel, ventilation controllers, or other sensors under evaluation. Data were summarized by The Maschhoff's research personnel into daily measurements and communicated to researchers at the University of Missouri (Columbia, Missouri, USA) for further summary, calculation, and analysis. All dataset manipulation, summarization, and calculation of additional variables were performed using the 'base' (R Core Team, 2023), 'stats' (R Core Team, 2023), and collection of 'tidyverse' (v2.0.0) (Wickham et al., 2019) packages in R (v4.3.1).

Metadata. The site, room within site, group within room within site, calendar date, day of week (**DOW**), number of days post placement (**DAYS**), and the total number of pigs (**INV**) was provided for each daily room measurement. Days post placement was defined as the number of days since the start of allotment. For example, a value of "0" for days post placement would indicate the first day a new group of pigs were moved into a room, while a value of "100" would indicate the 101st day on feed for a given group of pigs in a room. The total number of pigs for a room on a given day was recorded using the following formula:

$$\boldsymbol{INV} = prev.day\ ending\ inventory + today's\ additions - today's\ subtractions - today's\ mortalities$$

In addition, each day was categorized into one of four growth periods (**GP**) based on days post placement, as specified by standard analysis procedures within The Maschhoff's

production system: Early (0 to 41 days post placement; **GP_E**), Early Middle (42 to 83; **GP_EM**), Late Middle (84 to 125; **GP_EM**), and Late (126 and above; **GP_L**).

Mortality Measures. The number of pigs found dead (**DEAD**) and euthanized (**EUTH**) was recorded by barn personnel at the end of each workday. Total mortality was calculated as the sum of the daily number of found dead and euthanized pigs (**TMORT**). Each of the above values were divided by the total number of pigs in the room at the start of the day (INV) to calculate proportion of found dead (**PDEAD**), euthanized (**PEUTH**), and total mortality (**PTMORT**).

Medication Usage Measurements. The number of administered injectables was recorded by barn personnel each day for the following injectable classes: ceftiofur (**CEFT**), dexamethasone (**DEX**), enrofloxacin (**ENRO**), lincomycin (**LINCO**), penicillin (**PEN**), and tetracycline (**TETRA**). Any injectables administered that did not fall into any of the above classes were considered other (**OTHER**). Ceftiofur, enrofloxacin, and tetracycline injections were used primarily for treatment of respiratory infections and were summed to calculate the total daily primary respiratory injections (**PRIM**). Dexamethasone, lincomycin, and penicillin injections were used primarily for lameness or as secondary respiratory medications and were summed to calculate the total daily secondary respiratory injections (**SEC**). The total number of injections administered to pigs in a room on a given day (**TOTTR**) was calculated as follows:

$$\boldsymbol{TOTTR} = CEFT + DEX + ENRO + LINCO + PEN + TETRA + OTHER.$$

Each of the above measurements were divided by the total number of pigs in the room (INV) to yield the proportion of inventory treated on a given day (**PCEFT**, **PDEX**, **PENRO**, **PLINCO**, **PPEN**, **PTETRA**, **POTHER**, **PPRIM**, **PSEC**, and **PTOTTR**,

respectively). Lastly, the daily administration of broad usage water medications was also recorded by barn personnel and converted to a categorical variable (1 if administered, 0 if not; **WMED**).

Cough Measurements. Depending on physical dimension, each room was fit with three to five SoundTalks® (SoundTalks NV, Leuven, Belgium) sensors to detect the level of coughing in pig populations. Within each room, the daily values were averaged into one value that represented the respiratory health status (**REHS**) of a single room on a given day. These respiratory health status values ranged from 1 (very sick and lots of coughing) to 99 (healthy with minimal to no coughing). According to specifications set by SoundTalks® and Boehringer Ingelheim (Boehringer Ingelheim Vetmedica GmbH, Ingelheim am Rhein, Germany), a score greater than 60, less than 60 but greater than 40, and less than 40 was considered a green (**GREEN**), yellow (**YELLOW**), and red (**RED**) respiratory health status categorization, respectively, and dummy variables were created for each category as such.

Water Disappearance Measurements. Water disappearance measurements were recorded automatically using standard commercial ventilation controller systems (**VC**). Room water disappearance was recorded as the gallons of water disappeared since the previous day's observation (**WATER**). The number of gallons of water that disappeared from each room per pig (**WATER_HD**) per day was calculated using the following formula:

$$\boldsymbol{WATER\_HD} = \frac{WATER}{INV}$$

Temperature Measurements. Daily low, high, and setpoint temperatures were recorded automatically using standard commercial ventilation controller systems

(**LOW_TEMP**, **HIGH_TEMP**, and **SET_TEMP**, respectively). Using these measurements, the daily temperature range (**TEMP_RANGE**), low temperature set point deviation (**LOW_SET_DEV**), and high temperature set point deviation (**HIGH_SET_DEV**) were calculated with the following formulas:

$$\boldsymbol{TEMP\_RANGE} = HIGH\_TEMP - LOW\_TEMP,$$

$$\boldsymbol{LOW\_SET\_DEV} = LOW\_TEMP - SET\_TEMP,$$

$$and$$

$$\boldsymbol{HIGH\_SET\_DEV} = HIGH\_TEMP - SET\_TEMP.$$

Initial Dataset Preparation

Measurements described in the previous section were aggregated into a dataset containing 13,198 observations and 47 variables, and summary statistics for these variables are provided in Supplementary Table 4.2. Due to the wide range of distributions of the independent variables, specific rules were created for each variable to remove outliers with impossible biological values, and these rules are shown in Supplementary Table 4.3.

Definition of Mortality Outcomes

Two different mortality outcome variables were defined and included in subsequent machine learning prediction analyses, namely, mortality episode days (**MEP**) and extremely high mortality days (**EHMD**). The number of found dead pigs served as the basis for both mortality outcome variables (DEAD), as euthanization patterns are generally random and not representative of the true mortality patterns in a cohort.

Mortality Episode Days. Episodic mortality can be defined as sequences of days where the average daily mortality is higher than the population baseline at a given age for

a unique cohort of pigs reared on a single farm. In general, the identification of mortality episodes highlights acute periods when mortality risk is the most severe and removes random noise from a mortality time series consisting of only counts or rates. Using a novel algorithm developed by the authors in R (v4.3.1) and first proposed by Grohmann et al., mortality episodes were annotated in the current dataset as follows (Grohmann et al., 2024):

Applied to Entire Dataset:

Step 1. Split dataset into multiple smaller datasets by cohort.

Applied to each Split Dataset(s) from Step 1:

Step 2. Add leading and lagging zero-value observations to the DAYS and DEAD variables. The number of leading and lagging zeros can vary, but more than 1000 is not recommended, in general. This parameter is denoted as PADDING and was set to 625 in the current study.

Step 3. Estimate a penalized smoothing spline regression equation between DAYS and DEAD variables from Step 2 using the **smooth.spline** function of the 'stats' package in R. Select desired lambda value (i.e., smoothing parameter; parameter name: LAMBDA; set to 1×10^{-9} in the current study) for use in the **smooth.spline** function. Values of LAMBDA that are closer to zero are less restrictive and will produce a more flexible regression equation, similar to the behavior of higher order polynomial regression equations.

Step 4. Extract predicted DEAD values, defined as $\hat{y}_t$, from estimated smoothing spline regression equation. Calculate instantaneous slope, defined as $\frac{\Delta y}{\Delta x}$, for each observation using the following formula:

$$\frac{\Delta y}{\Delta x} = \frac{\hat{y}_t - \hat{y}_{t-1}}{DAYS_t - DAYS_{t-1}}$$

Step 5. Identify changepoints, which are days where the sign of the instantaneous slope is different from the previous day. Definitions of the two possible changepoints are described below:

3) Valley: Positive instantaneous slope on $DAYS_t$ and negative instantaneous slope on $DAYS_{t-1}$.
4) Peak: Negative instantaneous slope on $DAYS_t$ and positive instantaneous slope on $DAYS_{t-1}$.

Step 6. Identify mortality sequences, which are all days between each valley. For example, if a series of days has 3 valleys (A, B, and C), there would be 2 mortality sequences in the series (mortality sequence 1 = all days between valley A and B and mortality sequence 2 = all days between valley B and C).

Step 7. Within each mortality sequence, find the days with the maximum (i.e., fastest increasing) and minimum (i.e., fastest decreasing) instantaneous slopes. Consider the maximum and minimum instantaneous slopes in each mortality sequence the start and end of the given mortality sequence, respectively. Then, all days between the start and end day of the mortality sequence are considered a "first pass" mortality episode.

Final Classification of Mortality Episodes:

Step 8. Declassify first-pass mortality episodes that fail to meet the following criteria or minimum parameter values:

5) CRITERIA: The mortality episode contains a start, peak, and end.

6) MAGNITUDE PARAMETER: The average number of found dead pigs per day during the mortality episode must be greater than or equal to MAGNITUDE (set to 1 in the current study).
7) PROPORTION PARAMETER: The proportion of days during the mortality episode with at least 1 found dead pig must be greater than or equal to PROPORTION (set to 0.50 in the current study).
8) DURATION PARAMETER: The number of days in the mortality episode must not exceed DURATION (set to 21 in the current study).

Parameters in the above algorithm were set based on rigorous grid search experiments that evaluated the effect of each parameter on classified mortality episode duration, average found dead pigs per mortality episode day, and average number of mortality episodes per cohort of pigs. After this algorithm was applied to the current dataset, the variable mortality episode status was created, where all days between the identified start and end of the classified mortality episodes were considered "mortality episode days" and were coded as 1. All other days that were not within the start and end of a mortality episode were labeled "normal days" and coded as 0.

Extremely High Mortality Days. In addition to mortality episode status, extremely high mortality days were defined as singular days within a cohort where the number of found dead pigs exceeded expectation based on the baseline of the cohort and represented sporadic pig mortality. Cohort baseline for the number of found dead pigs was estimated using a negative binomial mixed regression model fit to the entire dataset with the 'glmmTMB' package (v1.1.8) (Brooks et al., 2017) of R (v4.3.1). The baseline model contained the fixed effects of growth period and calendar day of week and the random

effect of cohort and assumed an autoregressive order 1 covariance structure between residuals within a cohort. Predicted values were estimated for each observation and used in the following formula to classify each day as a normal or extremely high mortality day:

$$\boldsymbol{EHMD} = \frac{y - \hat{y} - M}{MAD} \geq 4,$$

where y is the number of found dead pigs for a room on a day, $\hat{y}$ is the predicted number of found dead pigs for a room on a day given the above model fit, M is the median of the residuals from the above model fit, and MAD is the median absolute deviation of the residuals from the above model fit. Given the highly zero-inflated and right skewed nature of mortality data (Varona and Sorensen, 2010), this method was utilized as a robust approximation of assigning *z*-scores using the mean and standard deviation. Any value of the above statistic that was greater than or equal to 4 was considered an extremely high mortality day and coded as 1, and all normal days were coded as 0. As an example of the behavior of these two methods for defining mortality outcomes, Figure 4.1 depicts actual classified mortality episodes and extremely high mortality days in an example selected cohort.

Data Preprocessing

A subset of the 47 variables was selected as feature variables (n = 21 features) for inclusion in machine learning prediction models based on relevance to the outcome and minimization of redundancy with other variables. For example, for antibiotic treatment variables, proportions were used instead of counts to adjust for varying numbers of pigs across farms due to barn dimensions and pig removals due to mortality. In addition, low and high temperature set point deviations were selected as opposed to recorded

temperatures, as these variables give a more accurate representation of the effect of external temperatures on internal temperatures within each room. Lastly, the numeric SoundTalks® respiratory health status was included over the color categorization variables to give a more precise indicator of the cough incidence within each room.

In the context of this study, both mortality outcomes are a daily recorded time series. Thus, to forecast future occurrence, lagged forms of each variable were created using the R package 'dplyr' (v1.1.3) (Wickham et al., 2023) from 3 to 5 days prior to the current observation (e.g, 3 days would be a feature variable recorded 3 days prior to the current observation). Observations recorded on day 0 hold little utility in a model built to forecast mortality outcomes, as the current observation could already lie within a mortality episode or be an extremely high mortality day. The additional removal of lags 1 to 2 ensured that models evaluated could predict each mortality outcome multiple days in advance to allow proper decisions concerning health intervention. After calculation of lagged features, there were 52 total features and 2 mortality outcomes available for further analysis.

Missing observations in any feature variable cause the removal of the respective observation from the dataset, in general. There are many methods to impute missing values, such as mean imputation or imputation based on predictive models that consist of variables present in all other variables. Because machine learning models to predict daily mortality outcomes in growing pigs have not been previously tested in the scientific literature, missing values were not imputed to preserve applicability in future studies. Instead, only complete observations were kept, and this resulted in a final dataset size of 5364 observations and 52 features.

Machine Learning Prediction Analysis

Three machine learning model frameworks were used to evaluate the predictive ability of each mortality outcome: 1) elastic net logistic regression (**ELNR**), 2) support vector machines (**SVM**), and 3) gradient boosted decision trees (i.e., XGBoost; **XGB**). Below, each of the three model frameworks are described in depth and in the context of the current study.

Elastic Net Logistic Regression. Logistic regression has been used extensively in classification analyses as baseline models (i.e., models in which more complex machine learning methods are expected to outperform), as these regression equations are computationally inexpensive and based on rigorous statistical theory (Wilson and Lorenz, 2015; Lasser et al., 2021). Standard logistic regression is based on link functions that vary linearly with the response variable; thus, multicollinearity issues frequently arise when independent variables are highly correlated with other independent variables. Elastic net logistic regression attempts to solve this issue by combining ridge (i.e., L2 regularization) and lasso (i.e., L1 regularization) penalization to shrink model coefficients towards 0 that are unimportant to the response variable, which induces an automatic feature selection process. In addition, elastic net logistic regression can select the most important features amongst groups of correlated independent variables (Zou and Hastie, 2005). To train an elastic net logistic regression model, two hyperparameters (i.e., parameters that affect the behavior of a machine learning model as opposed to individual features) were tested: 1) alpha, which was the mixing parameter between ridge and lasso regression, and 2) lambda, which was the shrinkage parameter. The package 'glmnet' v(4.1.8) (Tay et al., 2023) in R was used in the current study to fit all elastic net logistic

regression models. However, as a downfall, elastic net logistic regression requires explicit specification of potential interactions between or among feature variables, which motivates researchers to test more complicated machine learning models that automatically incorporate interactions amongst these features.

Gradient Boosted Decision Trees using XGBoost. Standard decision trees are a classification and regression method that builds a model resembling a tree using a sequence of if-then-else statements based on available features, which result in branches (i.e., split where a decision is made) and leaves (i.e., outcome of the branch split) (Navada et al., 2011; Joshi, 2023). A single decision tree is easy to interpret and generally a weak learner, which is a classifier that is prone to underfitting, simple, and hardly outperforms predictions based on random chance. Ensemble methods, such as gradient boosted decision trees [implemented using the 'xgboost' package (v2.0.3.1) (Chen et al., 2023) of R in the current study], attempt to solve the issue of simplicity through the fitting of many weak learners (i.e. standard decision trees). This process, referred to as boosting, first grows a decision tree from a random sample of the training data. Then, the algorithm uses samples from the first tree where performance was inadequate to grow the second tree with the goal of learning from previous classification errors. The sequential fitting of trees is repeated until a sufficiently "strong" learner is grown, and the weights from all sequential trees are combined to form one individual classifier (Joshi, 2023).

Although ensemble methods fix the problem of underfitting, these models can suffer from overfitting in the absence of properly selected hyperparameters. Thus, an extensive grid search consisting of the following 8 XGBoost hyperparameters (detailed in depth in Supplementary Table 4.6) was performed during model training: 1) eta: learning

rate, 2) max_depth: maximum tree depth, 3) gamma: minimum loss reduction, 4) subsample: sampling ratio of training observations, 5) colsample_bytree: sampling ratio of feature variables, 6) alpha: L1 regularization, 7) lambda: L2 regularization, and 8) max_delta_step: constraint parameter on learning rate (Chen et al., 2023). The optimal set of hyperparameters from the grid search were expected to provide maximum generalization to unforeseen, future data. For all XGBoost models, the negative log-likelihood (known as "logloss" in the 'xgboost' R package), was used as the evaluation metric during model training.

Support Vector Machines. An additional machine learning model framework, support vector machines, was tested in this study to complement the models described in previous sections. Support vector machine algorithms attempt to fit hyperplanes, which can be a linear or polynomial function, using the relationship between two or more features where the binary outcome is maximally separable. In this process, support vectors are first identified, which are data points that lie close to the margin between each class and are usually a small portion of the number of total samples in the training data set. After the support vectors are found, the linear or polynomial hyperplane is fit between the margin of support vectors of each outcome class, producing a boundary for classification of future observations (Joshi, 2023).

The R package "e1071" (v1.7.13) (Meyer et al., 2023) was used to fit support vector machine models during the training process in the current study. Linear, instead of polynomial, kernels were the basis of the hyperplanes to mitigate overfitting; thus, only one hyperparameter was evaluated, which was cost (*C*). Cost can be defined as a

regularization parameter that limits the influence of an individual sample and reduces the impact of noisy data, thus reducing overfitting in training datasets (Joshi, 2023).

Cross-Validation Approach. First, the entire data set was split into an initial training and holdout set (78% and 22%, respectively), which was comprised of 4176 training and 1188 holdout observations. Observations included in the holdout set were from the most recently marketed cohort of pigs from each room within farm. If the most recent cohort of pigs from a given room within farm was still in the growing process, that cohort was appended onto the previous complete cohort and included in the holdout dataset. Descriptive statistics for each variable in the training and holdout datasets are provided in Supplementary Tables 4.4 and 4.5, respectively. After the training and holdout split, all numerical and categorical features were mean and unit variance standardized and dummy encoded, respectively. Within the training set, validation folds were created using a "leave-one-out" approach based on farm, (e.g., A vs. B vs. C). Hyperparameters were trained for each model using a grid search (detailed in Supplementary Table 4.6) and were evaluated on the "left out" farm for each training fold. In addition to the model hyperparameters, a range of decision threshold values (0 to 1 by 0.01) were tested for each set of hyperparameters within each model framework (Supplementary Table 4.6). For balanced binary classes, 0.50 is typically chosen as the decision threshold, where predicted probabilities greater than 0.50 for an observation are considered a positive prediction. However, decision threshold-tuning, which moves the decision threshold closer to the prevalence (i.e., ratio of positive cases to total cases) of the minority class, has been shown to be an effective strategy in the remediation of severe class imbalances (Maloof, 2003; Zhi-Hua Zhou and Xu-Ying Liu, 2006). After the

optimal hyperparameters and thresholds were selected for each model, the performance of the classifier was tested on the holdout set to provide an estimation of actual forecasting and generalization ability. A depiction of the entire cross-validation approach is shown in Figure 2.2.

Model Evaluation and Selection. The number of true positive (**TP**), true negative (**TN**), false positive (**FP**), and false negative (**FN**) classifications for each set of hyperparameters within each model framework were summed for each validation fold. Then, each of these values were aggregated by like hyperparameters and summed across validation folds to yield overall values for the training portion of the cross-validation approach. An array of performance metrics using these values were calculated to identify optimal sets of hyperparameters for each model framework and are shown below:

$$True\ Positive\ Rate = Sensitivity = Recall = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = Specificity = \frac{TN}{TN + FP}$$

$$False\ Positive\ Rate = 1 - True\ Negative\ Rate = \frac{FP}{FP + TN}$$

$$False\ Negative\ Rate = 1 - True\ Positive\ Rate = \frac{FN}{FN + TP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Due to the highly imbalanced nature of each dependent variable (i.e., positive cases heavily outweigh negative cases), two variations of F-beta scores were also calculated to provide a more robust estimation of classifier performance than accuracy, which is biased towards 1 based on the prevalence of the minority class (Basha et al.,

2022). These metrics are weighted harmonic means of precision and recall and are shown below:

$$F1\ Score = \frac{2 \times Precision \times Recall}{2 \times Precision + Recall}$$

$$F0.5\ Score = \frac{1.25 \times Precision \times Recall}{0.25 \times Precision + Recall}$$

In general, a false positive diagnosis of a mortality outcome is more detrimental than a false negative outcome, due to the associated costs of wrongful intervention. Therefore, F0.5 score was chosen as the most important metric in selection of the optimal hyperparameters for each model framework during the training phase, as precision was weighted more highly than recall. In addition to these metrics, area under the receiver operator curve (**AUC**) and area under the precision recall curve (**AUCPR**) were also calculated using the 'MESS' package (v0.5.12) (Ekstrøm, 2023) in R. While AUC is more commonly used to assess binary classifiers, some studies have shown that AUCPR can be a more useful metric in imbalanced datasets due to the emphasis on the number of true positive predictions amongst all positive predictions, which was an important aspect of classifiers in the current study (Saito and Rehmsmeier, 2015; Sofaer et al., 2018). All metrics were calculated for each validation fold and on the final holdout set. Because there were 6 validation folds, mean and standard deviation of each metric were estimated for each set of hyperparameters within model framework during the training phase.

All performance metrics in this analysis were compared to a simple rule-based "baseline" classifier, a method commonly used to assess models trained to predict imbalanced class outcomes (Figueroa et al., 2017). The baseline predictions were established as follows:

$$MEP\ and\ EHMD = \begin{cases} 1\ if\ Early\ OR\ Early\ Middle\ Growth\ Period \\ 0\ otherwise \end{cases}$$

Models were considered informative if performance exceeded metrics estimated from the baseline classifier. In addition, model performance during the training phase and on the holdout dataset, as well as across mortality outcomes, was compared relative to baselines established in each dataset and for each outcome (Supplementary Table 4.9), which removed biases due to class imbalance differences. The above baseline classifier was used to evaluate all metrics except accuracy, which was compared against a naïve classifier that chose the majority class for each prediction, as illustrated by Basha et al. (2022).

RESULTS AND DISCUSSION

Prevalence of Mortality Episode and Extremely High Mortality Days

Daily measurements of mortality are highly variable and zero inflated, which means on most days, farmers will observe zero dead pigs. Variation and the degree of zero-inflation in mortality rate across cohorts of wean-to-finish pigs can be affected by a range of factors such as infectious diseases, genetics, sow farm attributes, climatic conditions, and others (Gebhardt et al., 2020a; Gebhardt et al., 2020b). Across the entire dataset in the current study, there were 8,558 observations with zero found dead pigs (66.3% of the total number of observations with a record). The distributional nature of the number of found dead pigs was reflected in the prevalence of positive cases for mortality episode days and extremely high mortality days in the dataset prepared for machine learning analyses (Table 4.1). Overall, only 14.6% and 8.9% of the total observations in the prepared dataset were classified as mortality episode and extremely high mortality days, respectively (Table 4.1), representing a considerable imbalance between classes of

each outcome variable. There was a further disparity in this class imbalance between the training (13.6 and 8.2% prevalence for MEP and EHMD, respectively) and holdout (18.0 and 11.4% prevalence for MEP and EHMD, respectively) datasets, as the prevalence for each mortality outcome was lower in the training dataset than the holdout dataset (Table 4.1).

Other factors such as age of the pigs and rearing location contribute to variation in the prevalence of each mortality outcome. Within the first few weeks after weaning, pigs are exposed to several stressors that can increase the likelihood of mortality, such as the distance traveled from sow to wean-to-finish farm, transition from milk to grain-based rations, mixing with pigs from multiple sow farms, independence from maternal antibodies, and increased environmental variability compared to sow farms (Le Dividich and Herpin, 1994; Campbell et al., 2013; Jayaraman and Nyachoti, 2017). In this study, mortality rate was higher and more variable in the early (GP_E) and early middle (GP_EM) growth periods relative to the late middle (GP_LM) and late (GP_L) growth periods, in general (Figure 4.3). Therefore, there were large differences in the prevalence of mortality outcomes in as pigs grew across the wean-to-market period. Mortality outcome prevalence in the early growth period was 35.0 and 14.5% in the training dataset and 34.7 and 19.6% in the holdout dataset for mortality episode and extremely high mortality days, respectively (Table 1). In the early middle growth period, positive cases were approximately half as prevalent in both datasets compared to the early growth period [18.4 and 10.1% (training) and 17.4 and 11.1% (holdout) for MEP and EHMD, respectively; Table 4.1]. However, the maximum prevalence in the late middle and late

growth periods was 1.6 and 5.2% in the training data set and 0.0 and 3.4% for mortality episode and extremely high mortality days, respectively (Table 4.1).

Class imbalances can have detrimental impacts on the performance of machine learning classifiers if handled improperly. As machine learning models use past data to learn patterns that accurately predict future outcomes, such models are easily biased towards prediction of the majority class, which causes suboptimal prediction performance (Guo et al., 2008). In addition, in uncontrolled settings such as commercial pig barns, the distributional nature of features and outcomes is rarely similar across subsequent data collections, and this phenomenon is defined as data drift (Moreno-Torres et al., 2012). In the current dataset, class imbalance and data drift were observed in the mortality outcomes. However, the main goal of any machine learning classifier is the optimization between bias (i.e. underfitting; overly simple models) and variance (i.e., overfitting; overly complex models) of future predictions, yielding a model with high generalization capability that is robust to changes in unforeseen data.

Both class imbalance and data drift were taken into consideration during the evaluation of each machine learning model framework through a custom cross-validation approach. Inherently, differences in mortality outcome prevalence were observed due to farm and growth period and the interaction of these two factors (Table 4.1). For example, in the training set, farm A reported a 30.9% mortality episode day prevalence in the early middle growth period, while farm B reported zero mortality episode days during the same growth phase (Table 4.1). Thus, to account for these biases, farm-specific validation folds were used during the training process to optimize performance across a range of underlying biological, management, and environmental scenarios (Figure 4.2). In

addition, intricate and comprehensive hyperparameter grid searches were tested for each model framework on each validation fold (Supplementary Table 4.6) that included a range of decision thresholds, which minimized the likelihood of selecting sets of hyperparameters that were suboptimal (i.e., a local minima) and identified a more refined decision threshold for each classifier. Lastly, to reduce the occurrence of false positive predictions due to differences in mortality outcome prevalence between early and late growth periods, F0.5 score was chosen as the most important metric for hyperparameter selection during the training phase, which placed the highest penalty on false positive predictions compared to F1 score, accuracy, AUC, and AUCPR.

The cross-validation approach and model evaluation was modeled based on commercial production scenarios and the forecasting of mortality outcomes during the rearing of future pig cohorts. Results from the machine learning analysis and considerations for use in pig production applications are discussed in the following sections.

Model Performance Across Farms

A 6-fold cross validation, where folds were specified based on farm, was used to select optimal hyperparameters (see Supplementary Table 4.10 for selected hyperparameters) and decision thresholds for each machine learning model. In general, k-fold cross validation yields reliable and unbiased performance metrics for a model's ability to predict outcomes in unforeseen data and provides a method to assess the generalization ability of a classifier (Kaariainen, 2006; Vabalas et al., 2019; Nti et al., 2021). Only results from the selected set of hyperparameters for each model are reported in this paper. Means and standard deviations of performance metrics for each mortality

outcome across farm are presented in Table 4.2. Individual performance metrics for prediction of mortality episode days and extremely high mortality days in each validation fold are shown in Supplementary Tables 4.7 and 4.8, respectively.

Each model framework reported high specificity (i.e., true negative rate) in the prediction of mortality episode days (0.82 ± 0.06, 0.83 ± 0.05, and 0.92 ± 0.02 for ELNR, SVM, and XGB, respectively; Table 4.2) and extremely high mortality days (0.95 ± 0.03, 0.92 ± 0.07, and 0.95 ± 0.04 for ELNR, SVM, and XGB, respectively; Table 4.2), which was mostly due to the low prevalence of these outcomes. Therefore, only metrics that did not consider specificity were considered in further discussion. As mentioned previously, F0.5 score (baseline = 0.30 and 0.14 for MEP and EHMD, respectively; Supplementary Table 4.9) was used as the metric to select hyperparameters. Overall, in the prediction of mortality episode days, XGBoost achieved a considerably higher F0.5 score (0.51 ± 0.17; 0.21 points above baseline; Table 4.2) compared to elastic net logistic regression (0.35 ± 13; Table 4.2) and support vector machines (0.30 ± 0.12; Table 4.2), which both either failed to or hardly outperform the baseline classifier. Much of this disparity between XGBoost and other models can be attributed to the large increases in precision (i.e., positive predictive value; 0.51 ± 0.20, 0.32 ± 0.16, and 0.28 ± 0.14 for XGB, ELNR, and SVM, respectively; Table 4.2; baseline = 0.26; Supplementary Table 4.9) and decreases in false positive rate (0.08 ± 0.02, 0.32 ± 0.16, and 0.28 ± 0.14 for XGB, ELNR, and SVM, respectively; Table 4.2). However, elastic net logistic regression offered more favorable recall (i.e., true positive rate) than the other models in mortality episode day prediction (0.56 ± 0.22, 0.51 ± 0.10, and 0.43 ± 0.22 for ELNR, SVM, and XGB, respectively; Table 4.2), although the cross-validation estimate was imprecise compared to XGBoost.

The impact of minimizing false positive rate in mortality episode prediction models is of utmost importance. In the current study, XGBoost reported an 0.08 false positive rate, which corresponds to the probability of a "false alarm", and this rate was considerably lower than the other tested machine learning models (Table 4.2). In the case of a false positive prediction, a farmer wastes additional resources such as labor, medications, and time to mitigate a mortality episode that does not occur. A false negative prediction in current commercial wean-to-finish pig barns is less costly, as management protocols for maximizing health and productivity in swine have been optimized over several decades; thus, the false negative prediction would likely go largely unnoticed. Nevertheless, the goal of mortality episode forecasting is to accurately target periods where pigs are at the highest risk for mortality multiple days in advance. In this regard, XGBoost also achieved the highest precision (0.51) by at least 0.19 points (Table 4.2); thus, this machine learning model framework is preferable to others for precisely targeting mortality episodes while limiting false positive predictions.

Across all performance metrics, the prediction of extremely high mortality days was more difficult than prediction of mortality episode days. In particular, F0.5 scores were low and only 0.02 to 0.10 points better than the baseline classifier (Table 4.2; Supplementary Table 4.9). Of the models, however, XGBoost yielded the most favorable and precise F0.5 score (0.24 ± 0.04; Table 4.2) compared to elastic net regression (0.21 ± 0.07; Table 4.2) and support vector machines (0.16 ± 0.08; Table 4.2), which was similar to the classification of mortality episode days. While the false positive rate of these classifiers was favorable (0.04 to 0.08; Table 4.2), the models were unable to accurately predict positive cases (precision = 0.23 ± 0.11, 0.16 ± 0.08, and 0.25 ± 0.14; recall = 0.16

± 0.10, 0.17 ± 0.16, and 0.20 ± 0.17 for ELNR, SVM, and XGB, respectively: Table 4.2). The largest disparity in performance of a classifier across mortality outcomes came from XGBoost models (0.21 vs. 0.10 points above baseline for F0.5 score in MEP and EHMD, respectively; Table 4.2; Supplementary Table 4.9), which suggests that prediction of episodic mortality is more feasible than sporadic mortality. Framing an extreme mortality outcome as an episode rather than a day reveals the sequences of days where pigs are found dead at a consistent and intense rate (Figure 4.1). On the other hand, an extremely high mortality day outcome can be standalone (Figure 4.1). As such, outlier days that are not a part of an episodic mortality outbreak are considered in each model, and predictor variables are likely not informative of these days due to the increased randomness of the outcome.

The study of postweaning mortality in pigs generally considers individual pig mortality (Fix et al., 2010) or retrospective mortality rates of entire cohorts (Losinger et al., 1998; Maes et al., 2004; Agostini et al., 2014; Magalhães et al., 2024). To the authors' knowledge, few studies have analyzed episodic mortality (Krahn, 2018; Mehling et al., 2019); however, no studies have assessed models to predict mortality episodes. Other studies have evaluated the prediction of mortality outcomes in preweaning (Rahman et al., 2023) and nursery (Magalhaes et al., 2023) pigs. Rahman et al. reported a high correlation between observed and predicted values for litter preweaning mortality rate (0.89) in random forest models; however, the data was from one sow farm and was randomly split into training and holdout sets and did not consider temporal variation (Rahman et al., 2023). In addition, Magalhaes et al. evaluated the ability of support vector machines to classify mortality of future nursery cohorts into high (> 5%) or low (<

5%) nursery mortality (Magalhaes et al., 2023). The authors reported high precision (0.92) and moderately high sensitivity (0.62), but the models were trained to predict overall instead of daily mortality outcomes. Furthermore, Bono et al. developed a dynamic monitoring system based on generalized linear models for mortality rates in sows and piglets (Bono et al., 2014). This technique forecasted mortality rates weekly for sow and piglet measures and applied statistical process control to forecasting errors for detection of significant deviations, therein termed "alarms" (Bono et al., 2014). While this system is similar in concept to the early warning system in this study, it may become infeasible to capture complex interactions as data sources increase in number and diversity, which is not an issue in machine learning models such as XGBoost. Because of the differences between experimental design, definition of mortality outcomes, and production stages across these studies and related to the current study, comparisons of model performance to our models are difficult.

Model Performance Across Time

While k-fold cross-validation is a reliable method for obtaining estimates of model performance on unseen data (Blum et al., 1999), many of these evaluations occur in processes that do not possess a time-series component. For a final comparison of each model's ability to forecast mortality outcomes, the optimal models from the training phase were tested on the most recent complete and/or partial cohort from each farm (Figure 4.2). This final step was required to validate performance measures obtained during the training phase in a commercial scenario, where data from past cohorts is used to support the decision-making process in future wean-to-market cohorts.

Table 4.3 presents the performance of each machine learning classifier on mortality outcomes in the holdout set. Much like the cross-validation estimates, XGBoost generally outperformed the other classifiers in the most important metrics for the prediction of mortality episode days. For example, XGBoost achieved an F0.5 score of 0.40 (baseline classifier = 0.30; Supplementary Table 4.9), while both elastic net regression and support vector machines reported scores of 0.36 (Table 4.3). In addition, predictions for XGBoost yielded a higher precision (0.39; Table 4.3; baseline classifier = 0.25; Supplementary Table 4.9) compared to both elastic net logistic regression and support vector machines (0.32 and 0.31, respectively; Table 4.3). However, the largest difference between XGBoost and the other machine learning models was observed for false positive rate, which was 0.39, 0.41, and 0.14 for elastic net logistic regression, support vector machines, and XGBoost, respectively (Table 4.3). As mentioned in the previous section, this result demonstrates the ability of XGBoost as a conservative classifier in the prediction of episodic mortality. Performance measures for classification of extremely high mortality days were worse (Table 4.3), which was similar to the across farm metrics (Table 4.2). Surprisingly, elastic net logistic regression outperformed all other models for all metrics except false positive rate (Table 4.3). Support vector machine classifiers reported a 0.01 for false positive rate; however, this favorable score was coupled with a low F0.5 score (0.03), precision (0.10) and recall (0.01), which meant this method was overly conservative (Table 4.3) in predictions in future cohorts. Nevertheless, while elastic net logistic regression achieved the highest F0.5 score (0.21), this method only outperformed the baseline classifier by 0.03 points (Table 4.3; Supplementary Table 4.9). Other methods failed to reach the baseline F0.5 score (Table

4.3; baseline classifier = 0.18; Supplementary Table 4.9), confirming the previous assumption that prediction of mortality episode days is more feasible than extremely high mortality days.

The above metrics were calculated based on the optimal decision threshold selected during hyperparameter tuning. While these values are important, receiver operating and precision recall curves evaluate the overall ability of a model as an informative classifier across a range of decision thresholds (Saito and Rehmsmeier, 2015; Sofaer et al., 2018). In the current study, Figures 4.4 and 4.5 depict receiver operating and precision-recall curves and estimates for area under each curve, respectively, for each machine learning classifier and mortality outcome evaluated on the holdout dataset. Based on suggestions by Swets, an AUC ≤ 0.50, 0.50 to 0.70, 0.70 to 0.90, and AUC ≥ 0.90 was considered indicative of a non-informative, weakly accurate, accurate, and highly accurate classifier, respectively (Swets, 1988; Greiner et al., 2000; Kavlak et al., 2023). There were negligible differences between each model for AUC in the prediction of mortality episode days (0.78, 0.77, and 0.77 for ELNR, SVM, and XGB, respectively; Figure 4.4), but these values suggest each classifier is accurate. However, XGBoost showed the highest early retrieval [i.e., fastest increase in the ROC curve from the origin; (Saito and Rehmsmeier, 2015)] of all algorithms. Area under the ROC curve was considerably less for prediction of extremely high mortality days, as expected, with elastic net logistic regression and XGBoost achieving the highest AUC values (0.67 and 0.63, respectively; Figure 4.4). Support vector machines were considered a non-informative classifier for extremely high mortality days in future pig cohorts (AUC = 0.50; Figure 4.4).

The major downfall in the use of receiver operating curves to evaluate classifier performance in imbalanced outcomes is the inclusion of specificity in the method, which is inherently inflated due to low outcome prevalence (Sofaer et al., 2018). On the other hand, precision-recall curves are not dependent on specificity, and thus, are generally more appropriate for rare events (Sofaer et al., 2018). In these curves, random performance is equal to prevalence (0.18 and 0.11 for MEP and EHMD, respectively; Table 4.1); however, the age dependent classifier achieved a 0.25 and 0.15 precision for mortality episode and extremely high mortality days, respectively (Supplementary Table 4.9), and was considered baseline (Figure 4.5). XGBoost achieved the most favorable relationship between precision and recall in future pig cohorts compared to other models (AUCPR = 0.46, 0.40, and 0.34 in XGB, ELNR, and SVM, respectively; Figure 4.5; Table 4.3). However, all models outperformed the baseline classifier by at least 0.09 points for AUCPR (0.21, 0.15, and 0.09 for XGB, ELNR, and SVM, respectively; Figure 4.5; Table 4.3; baseline AUCPR = 0.25; Figure 4.5; Supplementary Table 4.9). Like with receiver operating curves, XGBoost had the best early retrieval, reporting the highest precision and recall in the first half of the precision-recall curve (Figure 4.5). While the curve for XGBoost is not perfect, these results suggest that this machine learning model framework is highly informative when expectations are adjusted to account for prevalence of mortality episode days. Precision-recall curves for classification of extremely high mortality days were poor, unsurprisingly. Only elastic net logistic regression outperformed the baseline classifier for extremely high mortality days (AUCPR = 0.18, 0.15, and 0.12 for ELNR, XGB, and SVM, respectively; Figure 4.5;

Table 4.3), further bolstering the assumption that more refined classifiers or features are required to accurately forecast this mortality outcome.

There was a difference between the performance estimates during cross-validation and holdout testing for each classifier. Across all metrics, models performed more favorably during cross-validation as opposed to holdout evaluation, except for AUCPR. For example, F0.5 scores for mortality episode day prediction using XGBoost were 0.51 (Table 4.2) and 0.40 (Table 4.3) in the training and holdout phase, respectively (baseline classifier = 0.30 in both datasets; Supplementary Table 4.9). Furthermore, precision for this model was higher during cross-validation than holdout evaluation (0.51 and 0.39, respectively for MEP; Tables 4.2 and 4.3; 0.25 and 0.14 above baseline, respectively; Supplementary Table 4.9). However, there was less of a difference in false positive rate between the two phases for XGBoost (0.08 and 0.14 during training and holdout, respectively; Tables 4.2 and 4.3). On the other hand, AUCPR was higher for XGBoost during the holdout compared to training phase (0.46 vs. 0.35, respectively; Tables 4.2 and 4.3). The reasons for these differences, defined as generalization gap (An et al., 2021), are hard to identify. Data drift, especially in classifiers trained on small datasets, can cause degradation in model performance over time, especially in time-series data (Arora et al., 2024). In this study, there were differences in mortality outcome prevalence between the training and holdout datasets (Table 4.1), which likely contributed to the generalization gap. Nevertheless, the generalization gap was not overly large in XGBoost models, especially for critically important metrics such as false positive rate (0.06; Tables 4.2 and 4.3).

Variable Importance Measures

As machine learning models become more complex, interpretability of the relationship between features and outcomes decreases, and XGBoost and other models tested in this study are not immune to this problem (sometimes referred to as "black-box" models) (Sagi and Rokach, 2021). However, estimation of the relative importance of individual features is possible through built-in metrics provided by the XGBoost algorithm (Chen et al., 2023; Wang et al., 2024). Relative feature performance of the top 10 most important variables (summed across lagged days 3 to 5) for prediction of mortality episode days in XGBoost models is presented in Figure 4.6. The metric used to assess relative importance of each feature is termed gain, which is defined as the as a feature's improvement in the loss function when added as a new branch in a decision tree. Therefore, a feature with a 0.25 gain would have contributed to 25% of a model's improvement in negative log-likelihood (i.e., the loss function in this study) across all trees in an XGBoost model and is more favorable than smaller gain values. Across all features in a given model, individual gain values can be summed to equal 1. However, these values give no indication to the directional relationship between features and outcomes or causal effects, which requires controlled experiments and/or generalized linear mixed models and was outside the scope of this study.

As expected, days post weaning had the largest relative gain (0.26; Figure 4.6), as younger pigs were more likely to die than older pigs (Figure 4.3) and was corroborated by Krahn in pigs raised in similar conditions (Krahn, 2018). Water disappearance per pig also reported a relatively high gain (0.17; Figure 4.6). Daily deviations in water disappearance were tested during cross-validation, but these measures did not provide

any increase in model performance. Thus, the true impact of water disappearance per pig may be hard to distinguish from age, although XGBoost models are generally highly robust to multicollinearity between features (Li et al., 2023). Other studies have reported an impact of variation in water disappearance on mortality outcomes. Krahn found a negative relationship between percent daily deviations in water disappearance and the probability of a mortality event [Odds Ratio = 0.99; *P*-value < 0.05; (Krahn, 2018)]. In addition, early warning systems have been proposed to alert farmers to potential decreases in water intake of pigs to alleviate future mortality and detrimental growth performance (Vranken and Berckmans, 2017). The results in the current study suggest that water monitoring systems in pig barns could be beneficial when incorporated in machine learning models to predict mortality outcomes.

Cough incidence (REHS) was also found to influence mortality episode predictions, with a relative gain of 0.08 (Figure 4.6). The SoundTalks® sensor evaluated in the current study provided a 1 (high cough incidence) to 99 (low cough incidence) score daily for each cohort. This sensor has been evaluated in other studies to assess pig health. In pigs free from clinical respiratory infections, Pessoa et al. found that decreased air quality was associated with higher cough incidence in weaning-to-finishing pigs (Pessoa et al., 2022). In addition, Clavijo et al. reported that SoundTalks® sensors were we able to accurately identify *Mycoplasma hyopneumoniae* infections in pigs, and Silva et al. observed decreases in respiratory health status reported by SoundTalks® as *Mycoplasma hyopneumoniae* infections progressed post inoculation (Clavijo et al., 2021; Silva et al., 2022; Lagua et al., 2024). Sensors to track aspects of pig health have increased in popularity over the past several years, and based on results from this study,

inclusion in models to predict mortality outcomes improves the accuracy of classifier (Figure 4.6). Administration of antibiotics and temperature variation 3 to 5 days prior to a mortality episode day were less influential on model predictions compared to other features (0.03 to 0.08 relative gain; Figure 4.6); however, these variables are still important and easily included in complex machine learning models such as XGBoost.

CONCLUSIONS

In this study, machine learning models were tested to predict real-time, daily mortality outcomes in commercial wean-to-finish pig barns. These models combined data from several different sources, including manually collected production data and measurements from automated sensor technology. The most complex model, XGBoost, was able to more precisely forecast mortality outcomes than the other simpler methods, elastic net logistic regression and support vector machines. XGBoost achieved performance metrics that exceeded baseline classifiers based on class imbalance to a high degree when tested across farms and on future cohorts of pigs. Most importantly, XGBoost proved to be a conservative classifier, yielding low false positive rates, which are costly when attempting to mitigate mortality episodes in pig production. In addition, the definition of mortality outcomes as episodic as opposed to sporadic enabled the most accurate classifications when tested on new farms and future pig cohorts. The number of days post weaning, water disappearance, and cough incidence assessed by SoundTalks® sensors were the highly influential in increasing the accuracy of predictions from XGBoost models. This study serves as a foundation for future research in real -time prediction of mortality in pigs and demonstrates the efficacy of machine learning methods when applied to commercial pig production data.

FIGURES

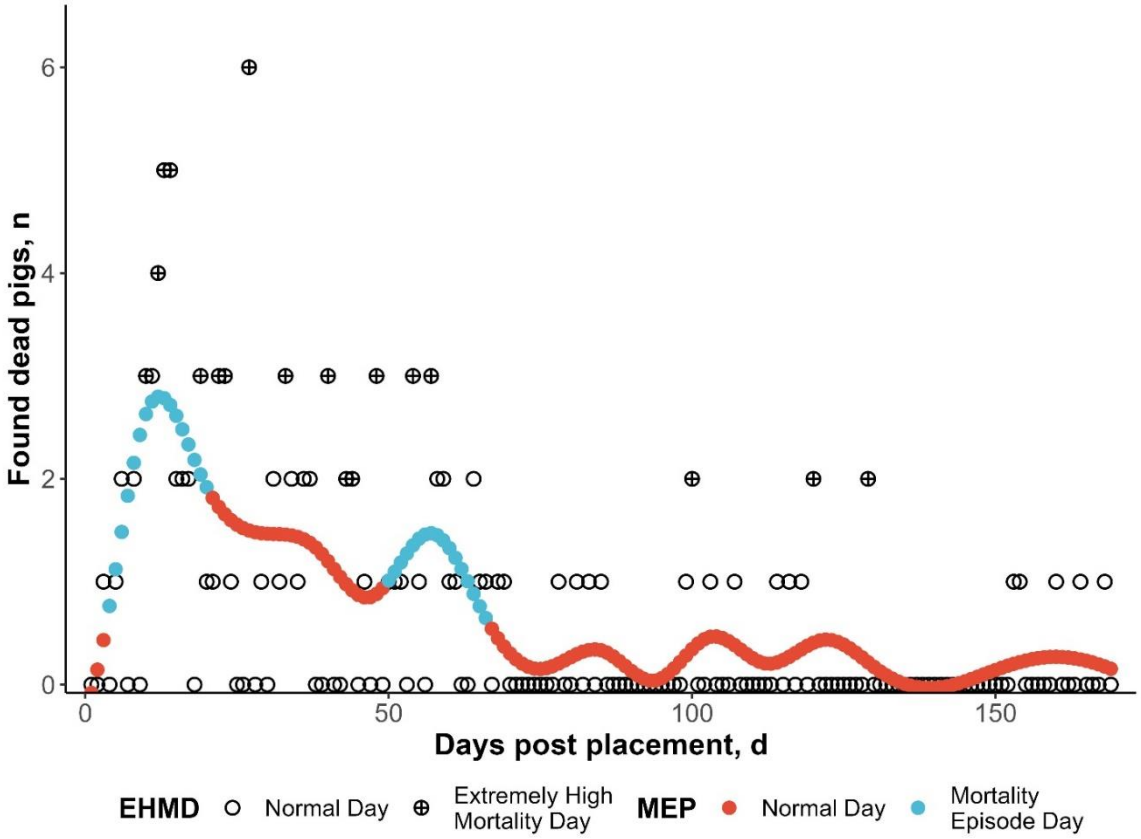


Figure 4.1. Behavior of defined mortality outcome variables, mortality episode and extremely high mortality days, in an example selected cohort of pigs. Colored points represent predicted values from smoothing spline regression equations used in the mortality episode classification algorithm. Black circles are observed values of found dead pigs per day.

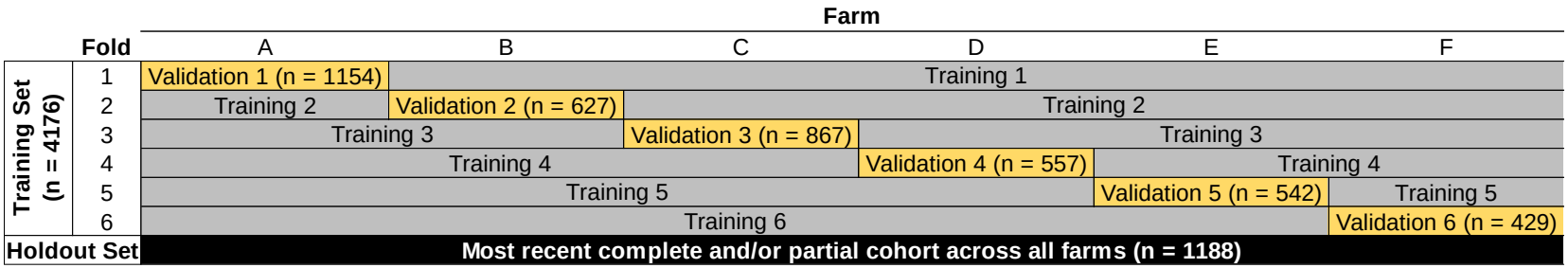


Figure 4.2. Cross validation scheme to select hyperparameters and evaluate prediction performance of machine learning models.

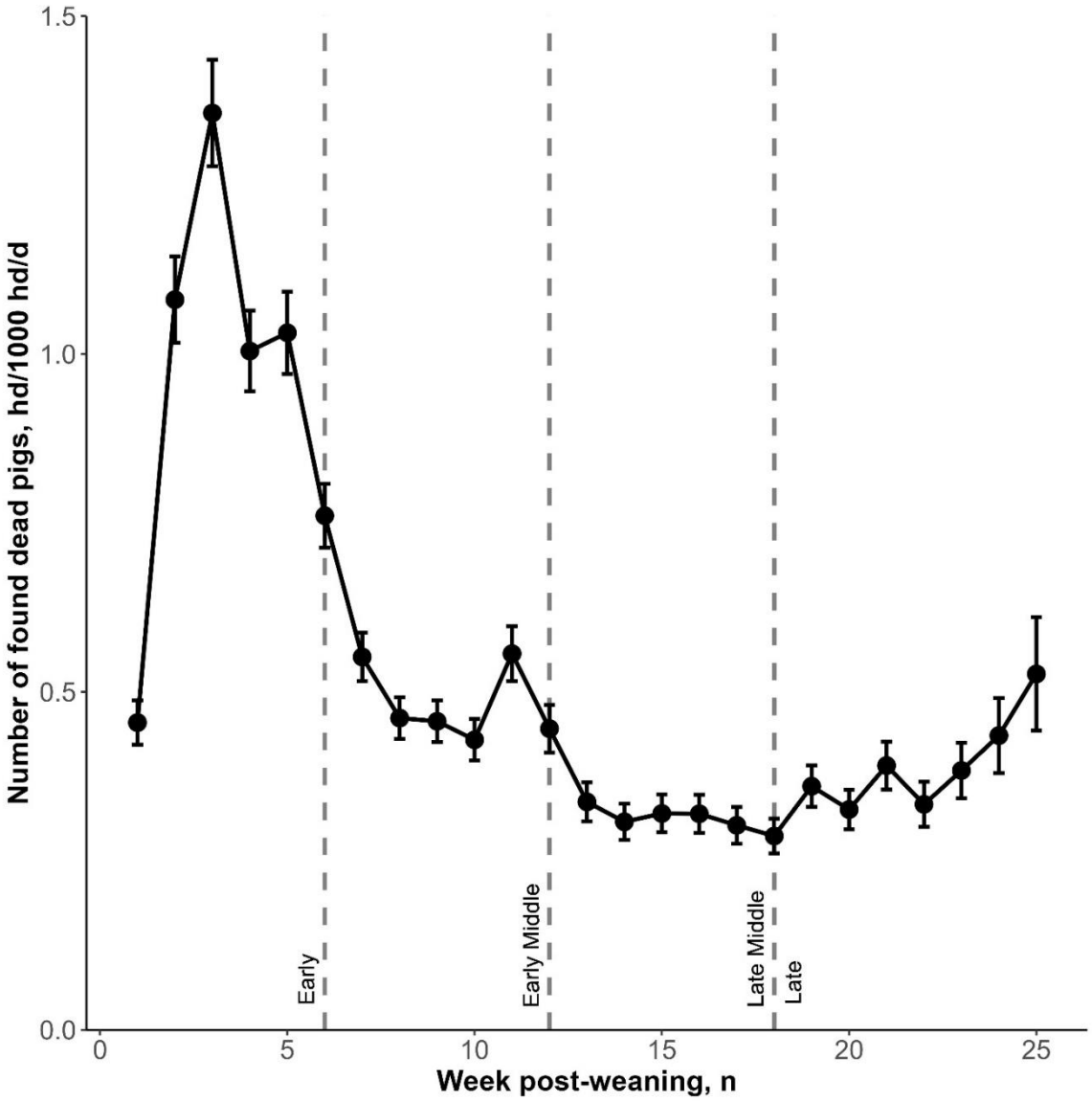


Figure 4.3. Relationship between weeks post-weaning and found dead pig rate across all cohorts in the current study. Vertical dashed lines delineate each growth period.

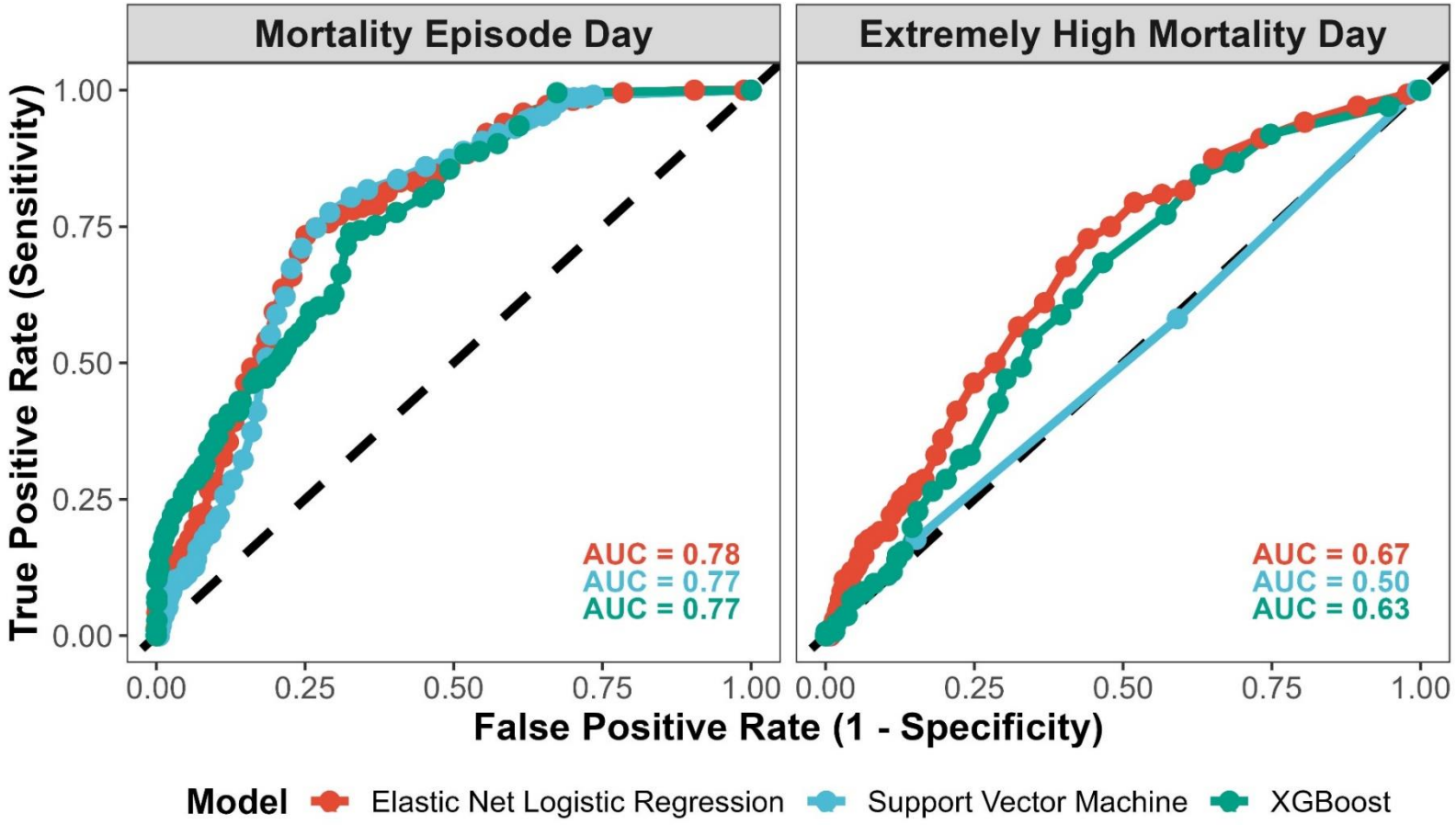


Figure 4.4. Receiver operating curves for elastic net logistic regression, support vector machine, and XGBoost models estimated from prediction performance for mortality outcomes on the holdout dataset.

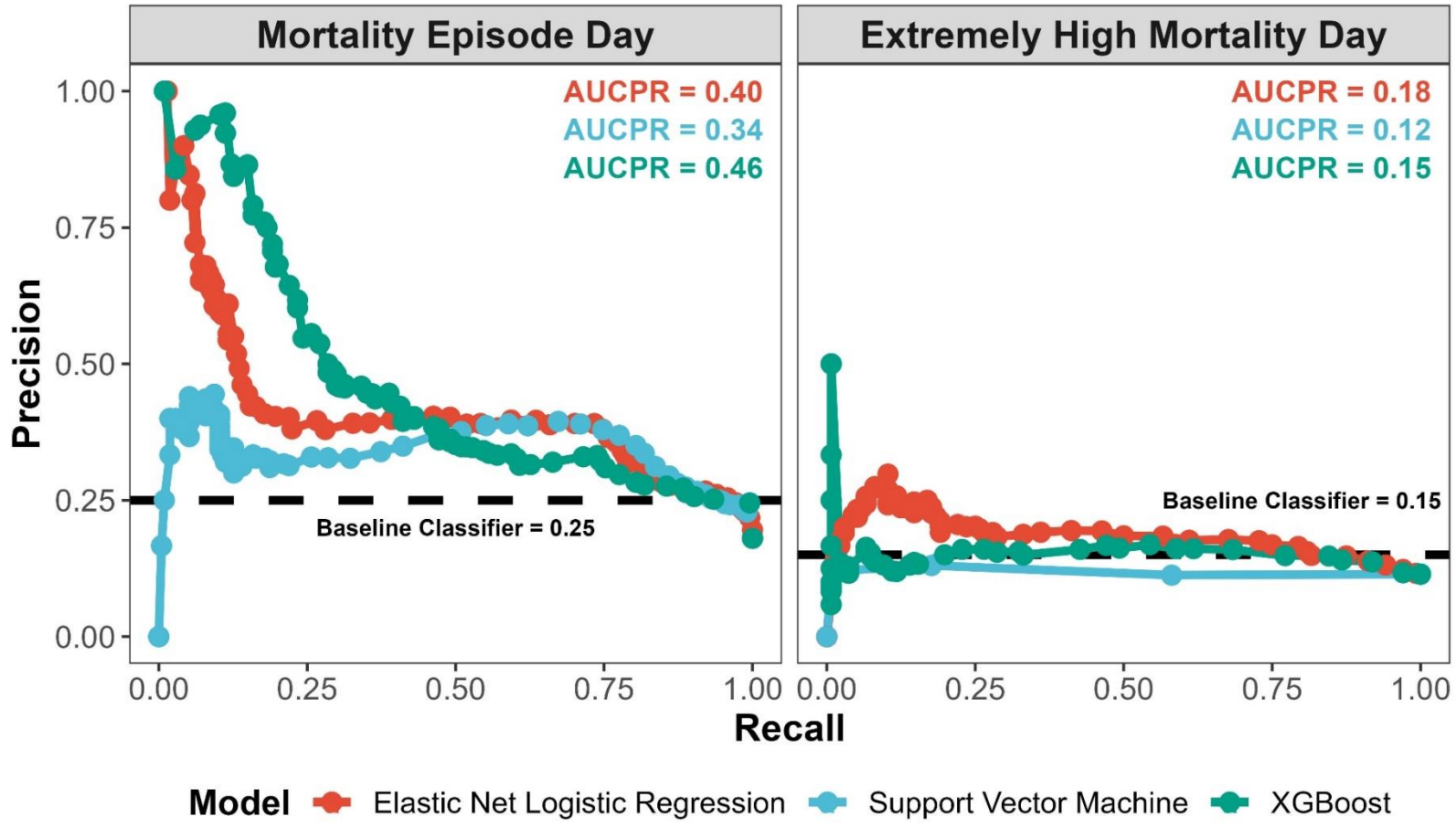


Figure 4.5. Precision-recall curves for elastic net logistic regression, support vector machine, and XGBoost models estimated from prediction performance for mortality outcomes on the holdout dataset.

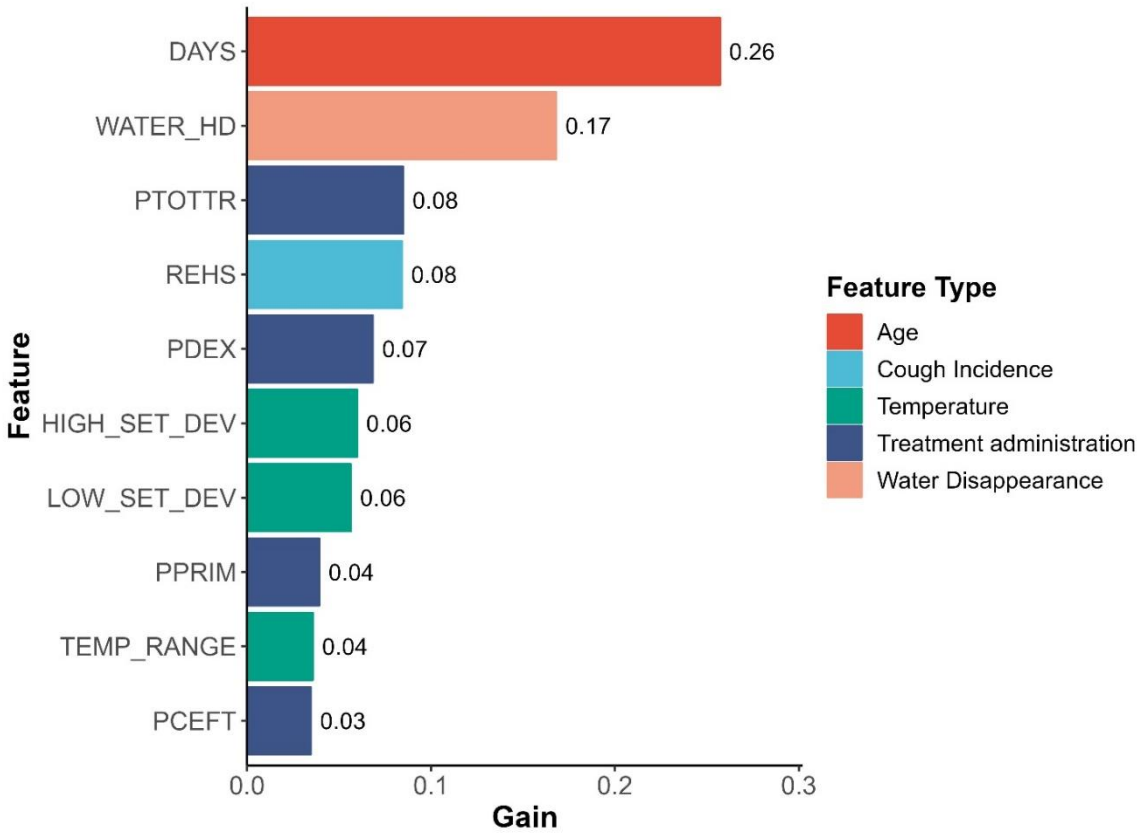


Figure 4.6. Relative variable importance, defined as gain, from the optimal XGBoost model during k-fold cross validation. DAYS = days post-placement; WATER_HD = water disappearance per pig (gal/hd/d); PTOTTR = number of total administered antibiotics as a proportion of room inventory; REHS = SoundTalks® cough incidence; PDEX = number of dexamethasone administrations as a proportion of room inventory; HIGH_SET_DEV = high temperature – set point temperature; LOW_SET_DEV = low temperature – set point temperature; PPRIM = number of primary respiratory injections as a proportion of room inventory; TEMP_RANGE = high temperature – low temperature; PCEFT = number of ceftiofur administrations as a proportion of room inventory.

TABLES

Table 4.1. Prevalence of mortality episode days and extremely high mortality days across farm, growth period, and training and holdout data sets.

	Training Set			Holdout Set		
Growth Period and Site	N	MEP, %[1]	EHMD, %[2]	N	MEP, %	EHMD, %
Early						
A	335	50.7	16.7	116	57.8	21.6
B	127	23.6	7.1	28	42.9	10.7
C	193	22.3	8.3	85	0.0	9.4
D	139	25.9	18.0	48	60.4	27.1
E	75	26.7	29.3	56	50.0	17.9
F	58	43.1	10.3	59	0.0	30.5
Total	**927**	**35.0**	**14.5**	**392**	**34.7**	**19.6**
Early Middle						
A	330	30.9	12.1	121	15.7	10.7
B	168	0.0	1.2	82	0.0	6.1
C	281	14.2	12.1	54	0.0	5.6
D	168	10.1	4.8	78	38.5	16.7
E	102	15.7	14.7	84	32.1	10.7
F	160	29.4	14.4	30	6.7	23.3
Total	**1209**	**18.4**	**10.1**	**449**	**17.4**	**11.1**
Late Middle						
A	250	0.0	0.8	84	0.0	1.2
B	168	0.0	1.8	64	0.0	6.3
C	177	0.0	3.4	0	-	-
D	161	0.0	3.7	32	0.0	0.0
E	127	0.0	9.4	9	0.0	0.0
F	77	5.2	2.6	41	0.0	0.0
Total	**960**	**0.4**	**3.2**	**230**	**0.0**	**2.2**
Late						
A	239	0.0	3.3	65	0.0	1.5
B	164	0.0	4.9	10	0.0	20.0
C	216	7.9	9.7	0	-	-
D	89	0.0	2.2	0	-	-
E	238	0.0	6.3	0	-	-
F	134	0.0	1.5	42	0.0	2.4
Total	**1080**	**1.6**	**5.2**	**117**	**0.0**	**3.4**
Grand Total	**4176**	**13.6**	**8.2**	**1188**	**18.0**	**11.4**

[1]MEP, % = mortality episode day prevalence

[2]EHMD, % = extremely high mortality day prevalence

Table 4.2. Optimal performance metrics for each model during hyperparameter tuning for classification of mortality episode and extremely high mortality days.

Model	Training Size, n	Cases, n	Threshold	F0.5 Score[1]	F1 Score[2]	Precision	Recall[3]	Spec.[4]	FPR[5]	FNR[6]	Accuracy	AUC[7]	AUCPR[8]
MEP													
ELNR	4176	567	0.17	0.35 ± 0.13	0.41 ± 0.08	0.32 ± 0.16	0.56 ± 0.22	0.82 ± 0.06	0.18 ± 0.06	0.44 ± 0.22	0.78 ± 0.03	0.84 ± 0.08	0.39 ± 0.16
SVM	4176	567	0.16	0.30 ± 0.12	0.34 ± 0.07	0.28 ± 0.14	0.43 ± 0.22	0.83 ± 0.05	0.17 ± 0.06	0.57 ± 0.22	0.77 ± 0.04	0.77 ± 0.10	0.28 ± 0.13
XGB	4176	567	0.35	0.51 ± 0.17	0.51 ± 0.12	0.51 ± 0.20	0.51 ± 0.10	0.92 ± 0.02	0.08 ± 0.02	0.49 ± 0.10	0.87 ± 0.02	0.83 ± 0.05	0.35 ± 0.14
EHMD													
ELNR	4176	343	0.19	0.21 ± 0.07	0.19 ± 0.07	0.23 ± 0.11	0.16 ± 0.10	0.95 ± 0.03	0.04 ± 0.03	0.84 ± 0.10	0.89 ± 0.04	0.65 ± 0.04	0.14 ± 0.05
SVM	4176	343	0.11	0.16 ± 0.08	0.16 ± 0.10	0.16 ± 0.08	0.17 ± 0.16	0.92 ± 0.07	0.08 ± 0.07	0.83 ± 0.16	0.86 ± 0.06	0.57 ± 0.10	0.10 ± 0.05
XGB	4176	343	0.20	0.24 ± 0.04	0.23 ± 0.05	0.25 ± 0.14	0.20 ± 0.17	0.95 ± 0.04	0.05 ± 0.04	0.79 ± 0.17	0.89 ± 0.04	0.65 ± 0.05	0.15 ± 0.07

[1]F0.5 Score = (1.25 × Precision × Recall) / (0.25 × Precision + Recall)
[2]F1 Score = (2 × Precision × Recall) / (Precision + Recall)
[3]Recall = True Positive Rate
[4]Specificity = True Negative Rate
[5]FPR = False Positive Rate
[6]FNR = False Negative Rate
[7]AUC = Area Under Receiver Operating Curve
[8]AUCPR = Area Under Precision Recall Curve

Table 4.3. Optimal performance metrics for each model evaluated on holdout set for classification of mortality episode and extremely high mortality days.

Model	Holdout Size, n	Cases, n	Threshold	F0.5 Score[1]	F1 Score[2]	Precision	Recall[3]	Specificity[4]	FPR[5]	FNR[6]	Accuracy	AUC[7]	AUCPR[8]
MEP													
ELNR	1188	214	0.17	0.36	0.45	0.32	0.81	0.61	0.39	0.19	0.65	0.78	0.40
SVM	1188	214	0.16	0.36	0.45	0.31	0.84	0.59	0.41	0.16	0.64	0.77	0.34
XGB	1188	214	0.35	0.40	0.40	0.39	0.41	0.86	0.14	0.59	0.78	0.77	0.46
EHMD													
ELNR	1188	136	0.19	0.21	0.24	0.19	0.33	0.81	0.19	0.67	0.76	0.67	0.18
SVM	1188	136	0.11	0.03	0.01	0.10	0.01	0.99	0.01	0.99	0.87	0.50	0.12
XGB	1188	136	0.20	0.14	0.14	0.13	0.15	0.87	0.13	0.85	0.79	0.63	0.15

[1]F0.5 Score = (1.25 × Precision × Recall) / (0.25 × Precision + Recall)
[2]F1 Score = (2 × Precision × Recall) / (Precision + Recall)
[3]Recall = True Positive Rate
[4]Specificity = True Negative Rate
[5]FPR = False Positive Rate
[6]FNR = False Negative Rate
[7]AUC = Area Under Receiver Operating Curve
[8]AUCPR = Area Under Precision Recall Curve

SUPPLEMENTARY TABLES

Supplementary Table 4.1. Description of each cohort of pigs in the current analysis.

Site	Room	Cohort	Daily observations, n	Start Date	End Date	Min. DPP[1]	Max. DPP[2]	SoundTalks®[3]
A	1	1	68	Friday, January 8, 2021	Tuesday, March 16, 2021	1	68	X
		2	56	Wednesday, March 31, 2021	Tuesday, May 25, 2021	1	56	X
		3	164	Thursday, June 10, 2021	Saturday, November 20, 2021	1	164	X
		4	168	Tuesday, November 30, 2021	Monday, May 16, 2022	1	168	X
		5	169	Monday, May 23, 2022	Monday, November 7, 2022	1	169	X
		6	181	Thursday, November 17, 2022	Tuesday, May 16, 2023	1	181	X
		7	129	Wednesday, June 7, 2023	Friday, October 13, 2023	1	129	X
	2	1	70	Tuesday, January 5, 2021	Monday, March 15, 2021	1	70	X
		2	66	Tuesday, March 23, 2021	Thursday, May 27, 2021	1	66	X
		3	171	Wednesday, June 2, 2021	Friday, November 19, 2021	1	171	X
		4	166	Thursday, December 2, 2021	Monday, May 16, 2022	1	166	X
		5	172	Monday, May 23, 2022	Thursday, November 10, 2022	1	172	X
		6	185	Monday, November 14, 2022	Wednesday, May 17, 2023	1	185	X
		7	138	Tuesday, May 30, 2023	Saturday, October 14, 2023	1	138	X
B	1	1	54	Sunday, July 31, 2022	Thursday, September 22, 2022	123	176	
		2	167	Tuesday, October 11, 2022	Sunday, March 26, 2023	1	167	
		3	178	Thursday, April 6, 2023	Saturday, September 30, 2023	1	178	
	2	1	46	Sunday, July 31, 2022	Wednesday, September 14, 2022	119	164	
		2	156	Thursday, October 20, 2022	Friday, March 24, 2023	1	156	
		3	167	Monday, April 17, 2023	Saturday, September 30, 2023	1	167	
	3	1	168	Thursday, March 18, 2021	Wednesday, September 1, 2021	1	168	X
		2	160	Thursday, September 30, 2021	Tuesday, March 8, 2022	1	160	X
		3	158	Thursday, April 7, 2022	Sunday, September 11, 2022	1	158	X
		4	159	Monday, October 17, 2022	Friday, March 24, 2023	1	159	X
		5	172	Wednesday, April 12, 2023	Saturday, September 30, 2023	1	172	X
	4	1	178	Monday, March 8, 2021	Wednesday, September 1, 2021	1	178	X
		2	169	Wednesday, September 22, 2021	Wednesday, March 9, 2022	1	169	X
		3	162	Monday, April 11, 2022	Monday, September 19, 2022	1	162	X
		4	151	Thursday, October 27, 2022	Sunday, March 26, 2023	1	151	X
		5	161	Sunday, April 23, 2023	Saturday, September 30, 2023	1	161	X
C	1	1	164	Saturday, December 5, 2020	Monday, May 17, 2021	1	164	X
		2	165	Thursday, June 10, 2021	Sunday, November 21, 2021	1	165	X
		3	175	Tuesday, December 14, 2021	Monday, June 6, 2022	1	175	X
		4	171	Tuesday, June 14, 2022	Thursday, December 1, 2022	1	171	X

		5	66	Wednesday, December 7, 2022	Friday, February 10, 2023	1	66	X
		6	73	Thursday, February 16, 2023	Saturday, April 29, 2023	1	73	X
		7	59	Monday, May 15, 2023	Wednesday, July 12, 2023	5	63	X
		8	68	Friday, July 21, 2023	Tuesday, September 26, 2023	1	68	X
		9	12	Tuesday, October 3, 2023	Saturday, October 14, 2023	1	12	X
		1	166	Thursday, December 3, 2020	Monday, May 17, 2021	1	166	X
		2	171	Friday, June 4, 2021	Sunday, November 21, 2021	1	171	X
		3	167	Wednesday, December 22, 2021	Monday, June 6, 2022	1	167	X
		4	161	Friday, June 24, 2022	Thursday, December 1, 2022	1	161	X
	2	5	65	Thursday, December 8, 2022	Friday, February 10, 2023	1	65	X
		6	66	Thursday, February 23, 2023	Saturday, April 29, 2023	8	73	X
		7	63	Thursday, May 11, 2023	Wednesday, July 12, 2023	1	63	X
		8	62	Thursday, July 27, 2023	Tuesday, September 26, 2023	7	68	X
		9	9	Friday, October 6, 2023	Saturday, October 14, 2023	4	12	X
		1	171	Tuesday, January 19, 2021	Thursday, July 8, 2021	1	171	X
		2	161	Saturday, July 31, 2021	Friday, January 7, 2022	1	161	X
	1	3	170	Tuesday, February 1, 2022	Wednesday, July 20, 2022	1	170	X
		4	169	Wednesday, August 10, 2022	Wednesday, January 25, 2023	1	169	X
		5	170	Saturday, February 18, 2023	Sunday, August 6, 2023	1	170	X
		6	34	Monday, August 28, 2023	Saturday, September 30, 2023	1	34	X
		1	164	Tuesday, January 26, 2021	Thursday, July 8, 2021	1	164	X
		2	160	Wednesday, August 4, 2021	Monday, January 10, 2022	1	160	X
	2	3	166	Saturday, February 5, 2022	Wednesday, July 20, 2022	1	166	X
D		4	160	Friday, August 19, 2022	Wednesday, January 25, 2023	1	160	X
		5	164	Friday, February 24, 2023	Sunday, August 6, 2023	1	164	X
		6	35	Sunday, August 27, 2023	Saturday, September 30, 2023	1	35	X
		1	25	Sunday, June 26, 2022	Wednesday, July 20, 2022	168	192	
		2	159	Saturday, August 20, 2022	Wednesday, January 25, 2023	1	159	
	3	3	163	Saturday, February 25, 2023	Sunday, August 6, 2023	1	163	
		4	33	Tuesday, August 29, 2023	Saturday, September 30, 2023	1	33	
		1	25	Sunday, June 26, 2022	Wednesday, July 20, 2022	153	177	
		2	169	Wednesday, August 10, 2022	Wednesday, January 25, 2023	1	169	
	4	3	178	Sunday, February 12, 2023	Tuesday, August 8, 2023	1	178	
		4	41	Monday, August 21, 2023	Saturday, September 30, 2023	1	41	
		1	95	Sunday, July 3, 2022	Wednesday, October 5, 2022	27	121	
	1	2	184	Tuesday, November 1, 2022	Wednesday, May 3, 2023	1	184	
E		3	131	Tuesday, May 23, 2023	Saturday, September 30, 2023	1	131	
	2	1	96	Sunday, July 3, 2022	Thursday, October 6, 2022	26	121	
		2	192	Monday, October 24, 2022	Wednesday, May 3, 2023	1	192	

		3	143	Thursday, May 11, 2023	Saturday, September 30, 2023	1	143	
		1	210	Thursday, January 28, 2021	Wednesday, August 25, 2021	1	210	X
		2	185	Tuesday, September 7, 2021	Thursday, March 10, 2022	1	185	X
	3	3	198	Tuesday, March 22, 2022	Wednesday, October 5, 2022	1	198	X
		4	197	Wednesday, October 19, 2022	Wednesday, May 3, 2023	1	197	X
		5	143	Thursday, May 11, 2023	Saturday, September 30, 2023	1	143	X
		1	203	Friday, February 5, 2021	Thursday, August 26, 2021	1	203	X
		2	185	Tuesday, September 7, 2021	Thursday, March 10, 2022	1	185	X
	4	3	192	Wednesday, March 30, 2022	Friday, October 7, 2022	1	192	X
		4	186	Thursday, October 27, 2022	Sunday, April 30, 2023	1	186	X
		5	136	Thursday, May 18, 2023	Saturday, September 30, 2023	1	136	X
		1	130	Tuesday, March 23, 2021	Friday, July 30, 2021	61	190	X
		2	178	Friday, August 6, 2021	Sunday, January 30, 2022	1	178	X
		3	172	Saturday, February 12, 2022	Tuesday, August 2, 2022	1	172	X
	1	4	78	Thursday, August 4, 2022	Thursday, October 20, 2022	1	78	X
		5	102	Friday, October 21, 2022	Monday, January 30, 2023	54	155	X
		6	177	Saturday, February 4, 2023	Sunday, July 30, 2023	1	177	X
		7	75	Tuesday, August 1, 2023	Saturday, October 14, 2023	1	75	X
F		1	126	Sunday, March 28, 2021	Saturday, July 31, 2021	66	191	X
		2	171	Saturday, August 14, 2021	Monday, January 31, 2022	1	171	X
		3	179	Saturday, February 5, 2022	Tuesday, August 2, 2022	1	179	X
	2	4	80	Friday, August 5, 2022	Sunday, October 23, 2022	1	80	X
		5	99	Monday, October 24, 2022	Monday, January 30, 2023	57	155	X
		6	171	Saturday, February 4, 2023	Monday, July 24, 2023	1	171	X
		7	75	Tuesday, August 1, 2023	Saturday, October 14, 2023	1	75	X

[1]Minimum days post-placement
[2]Maximum days post-placement
[3]Rooms with an "X" were outfitted with SoundTalks® sensors for the duration of the study

Supplementary Table 4.2. Descriptive statistics for entire dataset before data preprocessing for the machine learning prediction analysis.

Variable	Units	N	N, missing	Mean	SD[1]	Minimum	Median	Maximum
Inventory	pigs, n	12909	268	2028.8	980.02	1	1884	4317
Age								
Days post-weaning	d	13177	0	82.23	50.310	1	79	210
Early growth period	0 = All Others; 1 = Early	13177	0	0.275	0.4465	0	0	1
Early middle growth period	0 = All Others; 1 = Early Middle	13177	0	0.257	0.4372	0	0	1
Late middle growth period	0 =All Others; 1 = Late Middle	13177	0	0.230	0.4206	0	0	1
Late growth period	0 = All Others; 1 = Late	13177	0	0.238	0.4258	0	0	1
Mortality								
Count								
Found dead	pigs, n	12908	269	0.62	1.293	0	0	25
Euthanized	pigs, n	12908	269	0.47	1.649	0	0	50
Total found dead and euthanized	pigs, n	12908	269	1.09	2.519	0	0	60
Proportion								
Found dead	found dead pigs/inventory	12908	269	0.00033	0.000951	0	0	0.02298
Euthanized	euthanized pigs/inventory	12908	269	0.00028	0.001292	0	0	0.05591
Total found dead and euthanized	total found dead and euthanized pigs/inventory	12908	269	0.00061	0.001986	0	0	0.06607
Extremely High Mortality Day	0 = Normal Day; 1 = EHMD	12908	269	0.086	0.2803	0	0	1
Mortality Episode	0 = Normal Day; 1 = Episode Day	12207	970	0.130	0.3358	0	0	1
Medications								
Count								
Ceftiofur	injections/d, n	12156	1021	7.37	31.861	0	0	627
Dexamethasone	injections/d, n	12156	1021	4.76	17.278	0	0	306
Enrofloxacin	injections/d, n	12156	1021	6.29	22.769	0	0	700
Lincomycin	injections/d, n	12156	1021	6.04	20.277	0	0	422
Penicillin	injections/d, n	12156	1021	3.85	17.824	0	0	470
Tetracycline	injections/d, n	12156	1021	1.45	8.529	0	0	209
Other	injections/d, n	12156	1021	0.94	8.406	0	0	300
Primary respiratory	injections/d, n	12156	1021	15.11	38.903	0	0	700
Secondary respiratory	injections/d, n	12156	1021	14.65	32.460	0	0	470
Total treatments	injections/d, n	12156	1021	30.69	54.999	0	6	700
Proportion								
Ceftiofur	injections/pig/d, n	12156	1021	0.0024	0.00982	0	0	0.1985
Dexamethasone	injections/pig/d, n	12156	1021	0.0022	0.00790	0	0	0.1102
Enrofloxacin	injections/pig/d, n	12156	1021	0.0028	0.00951	0	0	0.1919
Lincomycin	injections/pig/d, n	12156	1021	0.0034	0.01064	0	0	0.2159
Penicillin	injections/pig/d, n	12156	1021	0.0014	0.00651	0	0	0.1907
Tetracycline	injections/pig/d, n	12156	1021	0.0006	0.00342	0	0	0.0761
Other	injections/pig/d, n	12156	1021	0.0005	0.00433	0	0	0.1102
Primary respiratory	injections/pig/d, n	12156	1021	0.0058	0.01355	0	0	0.1985
Secondary respiratory	injections/pig/d, n	12156	1021	0.0071	0.01517	0	0	0.2159
Total treatments	injections/pig/d, n	12156	1021	0.0134	0.02190	0	0.0036	0.2204

Water medications[2]	0 = No Water Medications; 1 = Water Medications	12143	1034	0.116	0.3204	0	0	1
Water disappearance								
Total water disappearance	gal	10329	2848	2179.1	1144.87	1.0	2195.0	9212.0
Water disappearance per pig	gal/pig	10329	2848	1.256	0.6634	0.101	1.306	4.000
Temperature								
Low temperature	°F	11317	1860	69.86	5.745	45.00	69.80	85.60
Set point temperature	°F	12750	427	68.80	6.479	60.00	67.20	89.00
High temperature	°F	11334	1843	79.17	7.505	45.40	79.00	100.70
Low temperature set-point deviation[3]	°F	11257	1920	1.09	4.561	-26.26	0.70	21.90
High temperature set-point deviation[4]	°F	11265	1912	10.41	8.007	-11.30	8.40	40.30
High - low temperature range	°F	11317	1860	9.33	5.717	0.00	8.00	38.90
Cough incidence								
SoundTalks® Respiratory Health Status	1 (high cough) to 99 (low cough) numeric score	9516	3661	83.91	18.626	1.33	92.50	99.00
Green Category	0 = All Others; 1 = Green Day	9516	3661	0.879	0.3263	0	1	1
Yellow Category	0 = All Others; 1 = Yellow Day	9516	3661	0.071	0.2573	0	0	1
Red Category	0 = All Others; 1 = Red Day	9516	3661	0.050	0.2178	0	0	1

[1]SD = Standard Deviation
[2]Observations coded "1" correspond to days where water medications were administered to a cohort of pigs
[3]Low temperature set point deviation = low temperature - set point temperature
[4]High temperature set point deviation = high temperature - set point temperature

Supplementary Table 4.3. Maximum and minimum removal threshold values for all variables.

Variable	Abbrev.	Minimum Threshold	Maximum Threshold	Comments
Inventory	INV	1	4500	If a day reported 0 inventory, all observations were converted to NA for that day
Age				
Days post-weaning	DAYS	1	-	-
Early growth period	GP_E	-	-	-
Early middle growth period	GP_EM	-	-	-
Late middle growth period	GP_LM	-	-	-
Late growth period	GP_L	-	-	-
Mortality				
Count				
Found dead	DEAD	0	-	Converted to NA if PTMORT was outside threshold range
Euthanized	EUTH	0	-	Converted to NA if PTMORT was outside threshold range
Total found dead and euthanized	TMORT	0	-	Converted to NA if PTMORT was outside threshold range
Proportion				
Found dead	PDEAD	0	-	Converted to NA if PTMORT was outside threshold range
Euthanized	PEUTH	0	-	Converted to NA if PTMORT was outside threshold range
Total found dead and euthanized	PTMORT	0	0.075	Converted to NA if PTMORT was outside threshold range
Extremely High Mortality Day	EHMD	-	-	-
Mortality Episode	MEP	-	-	-
Medications				
Count				
Ceftiofur	CEFT	0	-	Converted to NA if PTOTTR was outside threshold range
Dexamethasone	DEX	0	-	Converted to NA if PTOTTR was outside threshold range
Enrofloxacin	ENRO	0	-	Converted to NA if PTOTTR was outside threshold range
Lincomycin	LINCO	0	-	Converted to NA if PTOTTR was outside threshold range
Penicillin	PEN	0	-	Converted to NA if PTOTTR was outside threshold range
Tetracycline	TETRA	0	-	Converted to NA if PTOTTR was outside threshold range
Other	OTHER	0	-	Converted to NA if PTOTTR was outside threshold range
Primary respiratory	PRIM	0	-	Converted to NA if PTOTTR was outside threshold range
Secondary respiratory	SEC	0	-	Converted to NA if PTOTTR was outside threshold range
Total treatments	TOTTR	0	-	Converted to NA if PTOTTR was outside threshold range
Proportion				
Ceftiofur	PCEFT	0	-	Converted to NA if PTOTTR was outside threshold range
Dexamethasone	PDEX	0	-	Converted to NA if PTOTTR was outside threshold range
Enrofloxacin	PENRO	0	-	Converted to NA if PTOTTR was outside threshold range
Lincomycin	PLINCO	0	-	Converted to NA if PTOTTR was outside threshold range
Penicillin	PPEN	0	-	Converted to NA if PTOTTR was outside threshold range
Tetracycline	PTETRA	0	-	Converted to NA if PTOTTR was outside threshold range
Other	POTHER	0	-	Converted to NA if PTOTTR was outside threshold range
Primary respiratory	PPRIM	0	-	Converted to NA if PTOTTR was outside threshold range
Secondary respiratory	PSEC	0	-	Converted to NA if PTOTTR was outside threshold range
Total treatments	PTOTTR	0	0.25	Converted to NA if PTOTTR was outside threshold range

Water medications[2]	WMED	-	-	-
Water disappearance				
Total water disappearance	WATER	-	-	-
Water disappearance per pig	WATER_HD	0.1	7.5	-
Temperature				
Low temperature	LOW_TEMP	45	105	-
Set point temperature	SET_TEMP	60	90	-
High temperature	HIGH_TEMP	45	105	-
Low temperature set-point deviation[1]	LOW_SET_DEV	-	-	Converted to NA if LOW_TEMP was outside threshold range
High temperature set-point deviation[2]	HIGH_SET_DEV	-	-	Converted to NA if HIGH_TEMP was outside threshold range
High - low temperature range	TEMP_RANGE	-	-	Converted to NA if LOW_TEMP or HIGH_TEMP was outside threshold range
Cough incidence				
SoundTalks® Respiratory Health Status	REHS	1	99	-
Green Category	GREEN	-	-	-
Yellow Category	YELLOW	-	-	-
Red Category	RED	-	-	-

[1]Low temperature set point deviation = low temperature - set point temperature
[2]High temperature set point deviation = high temperature - set point temperature

Supplementary Table 4.4. Descriptive statistics for features and outcomes used in model training.

Variable	Units	N	Mean	SD[1]	Minimum	Median	Maximum
Outcomes							
Extremely High Mortality Day	0 = Normal Day; 1 = EHMD	4176	0.082	0.2746	0	0	1
Mortality Episode	0 = Normal Day; 1 = Episode Day	4176	0.136	0.3426	0	0	1
Features							
Age							
Days post-weaning	d	4176	86.65	47.384	7	82	191
Early growth period	0 = All Others; 1 = Early	4176	0.222	0.4156	0	0	1
Early middle growth period	0 = All Others; 1 = Early Middle	4176	0.290	0.4536	0	0	1
Late middle growth period	0 =All Others; 1 = Late Middle	4176	0.230	0.4208	0	0	1
Late growth period	0 = All Others; 1 = Late	4176	0.259	0.4379	0	0	1
Medications							
Ceftiofur	injections/pig/d, n	4176	0.0018	0.00716	0	0	0.0943
Dexamethasone	injections/pig/d, n	4176	0.0030	0.00900	0	0	0.1065
Enrofloxacin	injections/pig/d, n	4176	0.0028	0.00844	0	0	0.1665
Lincomycin	injections/pig/d, n	4176	0.0029	0.00909	0	0	0.1687
Penicillin	injections/pig/d, n	4176	0.0020	0.00595	0	0	0.0585
Tetracycline	injections/pig/d, n	4176	0.0005	0.00328	0	0	0.0761
Primary respiratory	injections/pig/d, n	4176	0.0051	0.01127	0	0	0.1665
Secondary respiratory	injections/pig/d, n	4176	0.0079	0.01456	0	0	0.1687
Total treatments	injections/pig/d, n	4176	0.0132	0.02034	0	0.0044	0.2143
Water medications[2]	0 = No Water Medications; 1 = Water Medications	4176	0.1159	0.32014	0	0	1
Water disappearance	gal/pig	4176	1.365	0.6214	0.156	1.447	3.965
Temperature							
Low temperature set-point deviation[3]	°F	4176	1.53	4.082	-18.60	1.30	14.60
High temperature set-point deviation[4]	°F	4176	10.03	7.370	-2.60	8.10	36.70
High - low temperature range	°F	4176	8.50	5.269	0.50	7.20	38.90
SoundTalks® Respiratory Health Status	1 (high cough) to 99 (low cough) numeric score	4176	87.65	16.763	10.00	95.00	99.00

[1]SD = Standard Deviation

[2]Observations coded "1" correspond to days where water medications were administered to a cohort of pigs

[3]Low temperature set point deviation = low temperature - set point temperature

[4]High temperature set point deviation = high temperature - set point temperature

Supplementary Table 4.5. Descriptive statistics for features and outcomes used in the holdout dataset.

Variable	Units	N	Mean	SD[1]	Minimum	Median	Maximum
Dependent Variables							
Extremely High Mortality Day	0 = Normal Day; 1 = EHMD	1188	0.114	0.3185	0	0	1
Mortality Episode	0 = Normal Day; 1 = Episode Day	1188	0.180	0.3845	0	0	1
Independent Variables							
Age							
Days post-weaning	d	1188	65.62	38.545	7	58.50	167
Early growth period	0 = All Others; 1 = Early	1188	0.330	0.4704	0	0	1
Early middle growth period	0 = All Others; 1 = Early Middle	1188	0.378	0.4851	0	0	1
Late middle growth period	0 =All Others; 1 = Late Middle	1188	0.194	0.3953	0	0	1
Late growth period	0 = All Others; 1 = Late	1188	0.098	0.2981	0	0	1
Medications							
Ceftiofur	injections/pig/d, n	1188	0.0029	0.01200	0	0	0.1513
Dexamethasone	injections/pig/d, n	1188	0.0031	0.00776	0	0	0.0490
Enrofloxacin	injections/pig/d, n	1188	0.0045	0.01227	0	0	0.1099
Lincomycin	injections/pig/d, n	1188	0.0037	0.01144	0	0	0.1253
Penicillin	injections/pig/d, n	1188	0.0042	0.01371	0	0	0.1907
Tetracycline	injections/pig/d, n	1188	0.0006	0.00299	0	0	0.0313
Primary respiratory	injections/pig/d, n	1188	0.0080	0.01639	0	0	0.1513
Secondary respiratory	injections/pig/d, n	1188	0.0110	0.01859	0	0.0009	0.1907
Total treatments	injections/pig/d, n	1188	0.0210	0.02629	0	0.0135	0.1907
Water medications[2]	0 = No Water Medications; 1 = Water Medications	1188	0.1835	0.38724	0	0	1
Water disappearance	gal/pig	1188	1.075	0.5114	0.149	1.092	3.181
Temperature							
Low temperature set-point deviation[3]	°F	1188	0.83	3.272	-16.70	0.50	14.10
High temperature set-point deviation[4]	°F	1188	8.23	6.614	-2.00	6.69	36.00
High - low temperature range	°F	1188	7.40	4.970	1.00	5.45	35.00
SoundTalks® Respiratory Health Status	1 (high cough) to 99 (low cough) numeric score	1188	74.61	22.320	10.60	82.67	99.00

[1]SD = Standard Deviation

[2]Observations coded "1" correspond to days where water medications were administered to a cohort of pigs

[3]Low temperature set point deviation = low temperature - set point temperature

[4]High temperature set point deviation = high temperature - set point temperature

Supplementary Table 4.6. Hyperparameters evaluated in a grid search experiment for each model framework during model training.

Model and Hyperparameter	R Package	Version[1]	Definition	Parameter Set
Elastic Net Logistic Regression				
alpha	glmnet	4.1.8	Elastic net mixing parameter. Alpha = 0 is ridge regression and alpha = 1 is lasso regression.	{0, …, 1} by 0.10
lambda			Regularization parameter. Larger values ($\lambda \geq 0$) correspond to greater shrinkage of regression coefficients.	{0, …, 1} by 0.10
decision threshold			Threshold for classification of positive cases	{0, …, 1} by 0.01
Support Vector Machines				
cost	e1071	1.7.13	Penalty factor that controls the tradeoff between errors of the model on training data and margin maximization. Small values can lead to underfitting while large values can lead to overfitting.	{1, …, 15} by 0.50
decision threshold			Threshold for classification of positive cases	{0, …, 1} by 0.01
XGBoost				
eta	xgboost	2.0.3.1	Learning rate or shrinkage factor of corrections from previously fitted models. Larger values reduce training time but can lead to overfitting.	{0.10, 0.30, 0.50, 0.70}
max_depth			Maximum depth of individual trees. Larger values lead to a more complex model.	{4, 6, 8, 10}
gamma			Minimum loss reduction required to make a further partition of a leaf node of a decision tree. Larger gamma values result in a more conservative algorithm.	{0, 2, 4, 6}
subsample			Sampling ratio of training observations prior to fitting decision trees.	{0.10, 0.25, 0.50, 0.75, 0.90, 1.00}
colsample_bytree			Sampling ratio of features (i.e., independent variables) when fitting a decision tree.	{0.10, 0.25, 0.50, 0.75, 0.90, 1.00}
alpha			L1 regularization term on model weights. Increasing this value results in a more conservative model.	{0, 1, 2, 5}
lambda			L2 regularization term on model weights. Increasing this value results in a more conservative model.	{0, 1, 2, 5}
max_delta_step			Constraint parameter on learning rate. Increasing this value results in more conservative models, reducing overfitting.	{0, 1, 2}
decision threshold			Threshold for classification of positive cases	{0, …, 1} by 0.01

[1]R (v4.3.1) package version used to fit models within each framework.

Supplementary Table 4.7. Performance metrics for each model by validation fold and overall for classification of mortality episode days.

Model	Fold Size, n	Cases, n	Threshold	F0.5 Score[1]	F1 Score[2]	Precision	Recall[3]	Specificity[4]	FPR[5]	FNR[6]	Accuracy	AUC[7]	AUCPR[8]
ELNR													
Farm													
A	1154	272	0.17	0.48	0.45	0.50	0.42	0.87	0.13	0.58	0.77	0.84	0.46
B	627	30	0.17	0.23	0.33	0.20	1.00	0.79	0.21	0.00	0.80	0.93	0.49
C	867	100	0.17	0.29	0.37	0.25	0.65	0.75	0.25	0.35	0.74	0.70	0.21
D	557	53	0.17	0.32	0.40	0.29	0.68	0.82	0.18	0.32	0.81	0.87	0.43
E	542	36	0.17	0.26	0.35	0.22	0.92	0.77	0.23	0.08	0.78	0.84	0.16
F	429	76	0.17	0.57	0.57	0.57	0.57	0.91	0.09	0.43	0.85	0.84	0.57
Overall	**4176**	**567**	**0.17**	**0.35 ± 0.13**	**0.41 ± 0.08**	**0.32 ± 0.16**	**0.56 ± 0.22**	**0.82 ± 0.06**	**0.18 ± 0.06**	**0.44 ± 0.22**	**0.78 ± 0.03**	**0.84 ± 0.08**	**0.39 ± 0.16**
SVM													
Farm													
A	1154	272	0.16	0.42	0.39	0.45	0.34	0.87	0.13	0.66	0.75	0.80	0.45
B	627	30	0.16	0.23	0.32	0.19	0.87	0.82	0.18	0.13	0.82	0.91	0.32
C	867	100	0.16	0.20	0.24	0.18	0.37	0.78	0.22	0.63	0.74	0.65	0.17
D	557	53	0.16	0.27	0.34	0.23	0.64	0.78	0.22	0.36	0.76	0.78	0.21
E	542	36	0.16	0.24	0.32	0.21	0.72	0.80	0.20	0.28	0.80	0.80	0.15
F	429	76	0.16	0.48	0.45	0.51	0.41	0.92	0.08	0.59	0.83	0.67	0.41
Overall	**4176**	**567**	**0.16**	**0.30 ± 0.12**	**0.34 ± 0.07**	**0.28 ± 0.14**	**0.43 ± 0.22**	**0.83 ± 0.05**	**0.17 ± 0.06**	**0.57 ± 0.22**	**0.77 ± 0.04**	**0.77 ± 0.10**	**0.28 ± 0.13**
XGB													
Farm													
A	1154	272	0.35	0.67	0.65	0.68	0.61	0.91	0.09	0.39	0.84	0.87	0.52
B	627	30	0.35	0.26	0.31	0.24	0.43	0.93	0.07	0.57	0.91	0.86	0.19
C	867	100	0.35	0.36	0.36	0.36	0.35	0.92	0.08	0.65	0.85	0.74	0.24
D	557	53	0.35	0.43	0.42	0.43	0.42	0.94	0.06	0.58	0.89	0.85	0.36
E	542	36	0.35	0.31	0.37	0.28	0.56	0.90	0.10	0.44	0.87	0.87	0.25
F	429	76	0.35	0.62	0.52	0.70	0.41	0.96	0.04	0.59	0.86	0.80	0.52
Overall	**4176**	**567**	**0.35**	**0.51 ± 0.17**	**0.51 ± 0.12**	**0.51 ± 0.20**	**0.51 ± 0.10**	**0.92 ± 0.02**	**0.08 ± 0.02**	**0.49 ± 0.10**	**0.87 ± 0.02**	**0.83 ± 0.05**	**0.35 ± 0.14**

[1]F0.5 Score = (1.25 × Precision × Recall) / (0.25 × Precision + Recall)
[2]F1 Score = (2 × Precision × Recall) / (Precision + Recall)
[3]Recall = True Positive Rate
[4]Specificity = True Negative Rate
[5]FPR = False Positive Rate
[6]FNR = False Negative Rate
[7]AUC = Area Under Receiver Operating Curve
[8]AUCPR = Area Under Precision Recall Curve

Supplementary Table 4.8. Performance metrics for each model by validation fold and overall for classification of extremely high mortality days.

Model	Fold Size, n	Cases, n	Threshold	F0.5 Score[1]	F1 Score[2]	Precision	Recall[3]	Specificity[4]	FPR[5]	FNR[6]	Accuracy	AUC[7]	AUCPR[8]
ELNR													
Farm													
A	1154	106	0.19	0.23	0.20	0.27	0.16	0.96	0.04	0.84	0.88	0.61	0.15
B	627	22	0.19	-	-	0.00	0.00	0.98	0.02	1.00	0.94	0.61	0.05
C	867	77	0.19	0.19	0.16	0.23	0.12	0.96	0.04	0.88	0.89	0.62	0.16
D	557	41	0.19	0.28	0.25	0.30	0.22	0.96	0.04	0.78	0.90	0.70	0.16
E	542	64	0.19	0.26	0.27	0.26	0.28	0.89	0.11	0.72	0.82	0.67	0.20
F	429	33	0.19	0.11	0.10	0.12	0.09	0.94	0.06	0.91	0.88	0.67	0.13
Overall	**4176**	**343**	**0.19**	**0.21 ± 0.07**	**0.19 ± 0.07**	**0.23 ± 0.11**	**0.16 ± 0.10**	**0.95 ± 0.03**	**0.04 ± 0.03**	**0.84 ± 0.10**	**0.89 ± 0.04**	**0.65 ± 0.04**	**0.14 ± 0.05**
SVM													
Farm													
A	1154	106	0.11	0.22	0.26	0.20	0.39	0.84	0.16	0.61	0.80	0.72	0.18
B	627	22	0.11	0.06	0.05	0.07	0.05	0.98	0.02	0.95	0.94	0.51	0.04
C	867	77	0.11	0.13	0.15	0.12	0.19	0.86	0.14	0.81	0.80	0.57	0.10
D	557	41	0.11	-	-	0.00	0.00	0.99	0.01	1.00	0.92	0.48	0.07
E	542	64	0.11	-	-	0.00	0.00	1.00	0.00	1.00	0.88	0.48	0.12
F	429	33	0.11	-	-	0.00	0.00	0.99	0.01	1.00	0.91	0.66	0.10
Overall	**4176**	**343**	**0.11**	**0.16 ± 0.08**	**0.16 ± 0.10**	**0.16 ± 0.08**	**0.17 ± 0.16**	**0.92 ± 0.07**	**0.08 ± 0.07**	**0.83 ± 0.16**	**0.86 ± 0.06**	**0.57 ± 0.10**	**0.10 ± 0.05**
XGB													
Farm													
A	1154	106	0.20	0.24	0.25	0.23	0.26	0.91	0.09	0.74	0.85	0.67	0.18
B	627	22	0.20	-	-	0.00	0.00	1.00	0.00	1.00	0.96	0.55	0.04
C	867	77	0.20	-	-	0.00	0.00	1.00	0.00	1.00	0.91	0.62	0.12
D	557	41	0.20	0.25	0.29	0.23	0.39	0.90	0.10	0.61	0.86	0.70	0.17
E	542	64	0.20	0.33	0.33	0.33	0.33	0.91	0.09	0.67	0.84	0.67	0.23
F	429	33	0.20	0.26	0.22	0.29	0.18	0.96	0.04	0.82	0.90	0.66	0.18
Overall	**4176**	**343**	**0.20**	**0.24 ± 0.04**	**0.23 ± 0.05**	**0.25 ± 0.14**	**0.20 ± 0.17**	**0.95 ± 0.04**	**0.05 ± 0.04**	**0.79 ± 0.17**	**0.89 ± 0.04**	**0.65 ± 0.05**	**0.15 ± 0.07**

[1]F0.5 Score = (1.25 × Precision × Recall) / (0.25 × Precision + Recall)
[2]F1 Score = (2 × Precision × Recall) / (Precision + Recall)
[3]Recall = True Positive Rate
[4]Specificity = True Negative Rate
[5]FPR = False Positive Rate
[6]FNR = False Negative Rate
[7]AUC = Area Under Receiver Operating Curve
[8]AUCPR = Area Under Precision Recall Curve

Supplementary Table 4.9. Performance metrics for baseline classifiers for each mortality outcome.

Dataset and Outcome	Fold Size, n	Cases, n	Threshold	F0.5 Score[1]	F1 Score[2]	Precision	Recall[3]	Specificity[4]	FPR[5]	FNR[6]	Accuracy	AUC[7]	AUCPR[8]
Training Dataset													
MEP	4176	567	-	0.30	0.40	0.26	0.96	0.56	0.44	0.04	0.86	0.50	0.25
EHMD	4176	343	-	0.14	0.21	0.12	0.75	0.51	0.49	0.25	0.92	0.50	0.09
Holdout Dataset													
MEP	1188	214	-	0.30	0.41	0.25	1.00	0.36	0.64	0.00	0.82	0.50	0.25
EHMD	1188	136	-	0.18	0.26	0.15	0.93	0.32	0.67	0.07	0.89	0.50	0.14

[1]F0.5 Score = (1.25 × Precision × Recall) / (0.25 × Precision + Recall)
[2]F1 Score = (2 × Precision × Recall) / (Precision + Recall)
[3]Recall = True Positive Rate
[4]Specificity = True Negative Rate
[5]FPR = False Positive Rate
[6]FNR = False Negative Rate
[7]AUC = Area Under Receiver Operating Curve
[8]AUCPR = Area Under Precision Recall Curve

Supplementary Table 4.10. Optimal hyperparameters for each mortality outcome and model framework based on F0.5 score evaluated on validation folds during model training.

Model and Hyperparameter	R Package	Version[1]	Definition	MEP	EHMD
Elastic Net Logistic Regression					
alpha	glmnet	4.1.8	Elastic net mixing parameter. Alpha = 0 is ridge regression and alpha = 1 is lasso regression.	0.10	0.00
lambda			Regularization parameter. Larger values ($\lambda \geq 0$) correspond to greater shrinkage of regression coefficients.	0.10	0.00
decision threshold			Threshold for classification of positive cases	0.17	0.19
Support Vector Machines					
cost	e1071	1.7.13	Penalty factor that controls the tradeoff between errors of the model on training data and margin maximization. Small values can lead to underfitting while large values can lead to overfitting.	12.5	2.0
decision threshold			Threshold for classification of positive cases	0.16	0.11
XGBoost					
eta	xgboost	2.0.3.1	Learning rate or shrinkage factor of corrections from previously fitted models. Larger values reduce training time but can lead to overfitting.	0.70	0.70
max_depth			Maximum depth of individual trees. Larger values lead to a more complex model.	6	6
gamma			Minimum loss reduction required to make a further partition of a leaf node of a decision tree. Larger gamma values result in a more conservative algorithm.	6	4
subsample			Sampling ratio of training observations prior to fitting decision trees.	1.00	0.10
colsample_bytree			Sampling ratio of features (i.e., independent variables) when fitting a decision tree.	0.75	1.00
alpha			L1 regularization term on model weights. Increasing this value results in a more conservative model.	1	5
lambda			L2 regularization term on model weights. Increasing this value results in a more conservative model.	1	1
max_delta_step			Constraint parameter on learning rate. Increasing this value results in more conservative models, reducing overfitting.	2	2
decision threshold			Threshold for classification of positive cases	0.35	0.20

[1]R (v4.3.1) package version used to fit models within each framework.

LITERATURE CITED

Agostini, P. S., A. G. Fahey, E. G. Manzanilla, J. V. O'Doherty, C. De Blas, and J. Gasa. 2014. Management factors affecting mortality, feed intake and feed conversion ratio of grow-finishing pigs. Animal. 8:1312–1318. doi:10.1017/S1751731113001912.

Ai, H., L. Huang, and J. Ren. 2013. Genetic Diversity, Linkage Disequilibrium and Selection Signatures in Chinese and Western Pigs Revealed by Genome-Wide SNP Markers. C. A. Kozak, editor. PLoS ONE. 8:e56001. doi:10.1371/journal.pone.0056001.

An, C., Y. W. Park, S. S. Ahn, K. Han, H. Kim, and S.-K. Lee. 2021. Radiomics machine learning study with a small sample size: Single random training-test set split may lead to unreliable results. K. N. Q. Le, editor. PLOS ONE. 16:e0256152. doi:10.1371/journal.pone.0256152.

Arora, S., R. Rani, and N. Saxena. 2024. A systematic review on detection and adaptation of concept drift in streaming data using machine learning techniques. WIREs Data Min. Knowl. Discov. e1536. doi:10.1002/widm.1536.

Barghi, N., J. Hermisson, and C. Schlötterer. 2020. Polygenic adaptation: a unifying framework to understand positive selection. Nat. Rev. Genet. 21:769–781. doi:10.1038/s41576-020-0250-z.

Basha, S. J., S. R. Madala, K. Vivek, E. S. Kumar, and T. Ammannamma. 2022. A Review on Imbalanced Data Classification Techniques. In: 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA). IEEE, Coimbatore, India. p. 1–6. Available from: https://ieeexplore.ieee.org/document/9753392/

Berckmans, D., and M. Guarino. 2017. From the Editors: Precision livestock farming for the global livestock sector. Anim. Front. 7:4–5. doi:10.2527/af.2017.0101.

Blum, A., A. Kalai, and J. Langford. 1999. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In: Proceedings of the twelfth annual conference on Computational learning theory. p. 203–209.

Bono, C., C. Cornou, S. Lundbye-Christensen, and A. Ringgaard Kristensen. 2014. Dynamic production monitoring in pig herds III. Modeling and monitoring mortality rate at herd level. Livest. Sci. 168:128–138. doi:10.1016/j.livsci.2014.08.003.

Brooks, M., et al. 2017. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. R J. 9:378. doi:10.32614/RJ-2017-066.

Campbell, J. M., J. D. Crenshaw, and J. Polo. 2013. The biological stress of early weaned piglets. J. Anim. Sci. Biotechnol. 4:19. doi:10.1186/2049-1891-4-19.

Carey, V. 2023. gwascat: representing and modeling data in the EMBL-EBI GWAS catalog. Available from: https://bioconductor.org/packages/gwascat

Chen, G. K., P. Marjoram, and J. D. Wall. 2008. Fast and flexible simulation of DNA sequence data. Genome Res. 19:136–142. doi:10.1101/gr.083634.108.

Chen, T., et al. 2023. xgboost: Extreme Gradient Boosting. Available from: https://github.com/dmlc/xgboost

Clavijo, M. J., et al. 2021. Mycoplasma hyopneumoniae Surveillance in Pig Populations: Establishing Sampling Guidelines for Detection in Growing Pigs. B. Fenwick, editor. J. Clin. Microbiol. 59:e03051-20. doi:10.1128/JCM.03051-20.

Conway, J. R., A. Lex, and N. Gehlenborg. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. J. Hancock, editor. Bioinformatics. 33:2938–2940. doi:10.1093/bioinformatics/btx364.

Cunningham, F., et al. 2022. Ensembl 2022. Nucleic Acids Res. 50:D988–D995. doi:10.1093/nar/gkab1049.

Decker, J. E. 2015. Agricultural Genomics: Commercial Applications Bring Increased Basic Research Power. G. Gibson, editor. PLOS Genet. 11:e1005621. doi:10.1371/journal.pgen.1005621.

Decker, J. E., et al. 2012. A novel analytical method, Birth Date Selection Mapping, detects response of the Angus (Bos taurus) genome to selection on complex traits. BMC Genomics. 13:606. doi:10.1186/1471-2164-13-606.

Dufrasne, M., I. Misztal, S. Tsuruta, N. Gengler, and K. A. Gray. 2014. Genetic analysis of pig survival up to commercial weight in a crossbred population. Livest. Sci. 167:19–24. doi:10.1016/j.livsci.2014.05.001.

Ekenberg, C., et al. 2019. Association Between Single-Nucleotide Polymorphisms in HLA Alleles and Human Immunodeficiency Virus Type 1 Viral Load in Demographically Diverse, Antiretroviral Therapy–Naive Participants From the Strategic Timing of AntiRetroviral Treatment Trial. J. Infect. Dis. 220:1325–1334. doi:10.1093/infdis/jiz294.

Ekstrøm, C. 2023. MESS: Miscellaneous Esoteric Statistical Scripts. Available from: https://CRAN.R-project.org/package=MESS

Ellis, M., J. P. Chadwick, W. C. Smith, and R. Laird. 1988. Index selection for improved growth and carcass characteristics in a population of Large White pigs. Anim. Sci. 46:265–275. doi:10.1017/S0003356100042331.

Figueroa, G., Y.-S. Chen, N. Avila, and C.-C. Chu. 2017. Improved practices in machine learning algorithms for NTL detection with imbalanced data. In: 2017 IEEE Power & Energy Society General Meeting. IEEE, Chicago, IL. p. 1–5. Available from: http://ieeexplore.ieee.org/document/8273852/

Fix, J. S., J. P. Cassady, J. W. Holl, W. O. Herring, M. S. Culbertson, and M. T. See. 2010. Effect of piglet birth weight on survival and quality of commercial market swine. Livest. Sci. 132:98–106. doi:10.1016/j.livsci.2010.05.007.

Fonseca, P. A. S., A. Suárez-Vega, G. Marras, and Á. Cánovas. 2020. GALLO: An R package for genomic annotation and integration of multiple data sources in livestock for positional candidate loci. GigaScience. 9:giaa149. doi:10.1093/gigascience/giaa149.

Gaynor, C., G. Gorjanc, and J. Hickey. 2020. AlphaSimR: Breeding Program Simulations. Available from: https://CRAN.R-project.org/package=AlphaSimR

Gaynor, R. C., G. Gorjanc, and J. M. Hickey. 2021. AlphaSimR: an R package for breeding program simulations. D.-J. De Koning, editor. G3 GenesGenomesGenetics. 11:jkaa017. doi:10.1093/g3journal/jkaa017.

Gebhardt, J. T., et al. 2020a. Postweaning mortality in commercial swine production. I: review of non-infectious contributing factors. Transl. Anim. Sci. 4:462–484. doi:10.1093/tas/txaa068.

Gebhardt, J. T., et al. 2020b. Postweaning mortality in commercial swine production II: review of infectious contributing factors. Transl. Anim. Sci. 4:485–506. doi:10.1093/tas/txaa052.

Gouveia, J. J. de S., M. V. G. B. da Silva, S. R. Paiva, and S. M. P. de Oliveira. 2014. Identification of selection signatures in livestock species. Genet. Mol. Biol. 37:330–342. doi:10.1590/S1415-47572014000300004.

Greiner, M., D. Pfeiffer, and R. D. Smith. 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev. Vet. Med. 45:23–41. doi:10.1016/S0167-5877(00)00115-X.

Grohmann, C. J., et al. 2024. 66 A novel analytical method for identifying periods of increased mortality in individual wean-to-finish pig barns. J. Anim. Sci. 102 (Supplement_2):37–38. doi:10.1093/jas/skae102.044.

Guo, X., Y. Yin, C. Dong, G. Yang, and G. Zhou. 2008. On the Class Imbalance Problem. In: 2008 Fourth International Conference on Natural Computation. IEEE, Jinan, Shandong, China. p. 192–201. Available from: http://ieeexplore.ieee.org/document/4667275/

Gurgul, A., et al. 2018. A genome-wide detection of selection signatures in conserved and commercial pig breeds maintained in Poland. BMC Genet. 19:95. doi:10.1186/s12863-018-0681-0.

Hazel, L. N. 1943. THE GENETIC BASIS FOR CONSTRUCTING SELECTION INDEXES. Genetics. 28:476–490. doi:10.1093/genetics/28.6.476.

Hazel, L. N., G. E. Dickerson, and A. E. Freeman. 1994. The Selection Index—Then, Now, and for the Future. J. Dairy Sci. 77:3236–3251. doi:10.3168/jds.S0022-0302(94)77265-9.

Hazel, L. N., and J. L. Lush. 1942. THE EFFICIENCY OF THREE METHODS OF SELECTION*. J. Hered. 33:393–399. doi:10.1093/oxfordjournals.jhered.a105102.

Hermisson, J., and P. S. Pennings. 2005. Soft Sweeps. Genetics. 169:2335–2352. doi:10.1534/genetics.104.036947.

Höllinger, I., P. S. Pennings, and J. Hermisson. 2019. Polygenic adaptation: From sweeps to subtle frequency shifts. J. C. Fay, editor. PLOS Genet. 15:e1008035. doi:10.1371/journal.pgen.1008035.

Hu, Z.-L., C. A. Park, and J. M. Reecy. 2022. Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. Nucleic Acids Res. 50:D956–D961. doi:10.1093/nar/gkab1116.

Ibáñez-Escriche, N., S. Forni, J. L. Noguera, and L. Varona. 2014. Genomic information in pig breeding: Science meets industry needs. Livest. Sci. 166:94–100. doi:10.1016/j.livsci.2014.05.020.

Iida, T., and M. A. Lilly. 2004. *missing oocyte* encodes a highly conserved nuclear protein required for the maintenance of the meiotic cycle and oocyte identity in *Drosophila*. Development. 131:1029–1039. doi:10.1242/dev.01001.

Jayaraman, B., and C. M. Nyachoti. 2017. Husbandry practices and gut health outcomes in weaned piglets: A review. Anim. Nutr. 3:205–211. doi:10.1016/j.aninu.2017.06.002.

Jonas, E., and D.-J. de Koning. 2015. Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. Front. Genet. 6. doi:10.3389/fgene.2015.00049. Available from: http://journal.frontiersin.org/Article/10.3389/fgene.2015.00049/abstract

Joshi, A. V. 2023. Machine Learning and Artificial Intelligence. Springer International Publishing, Cham. Available from: https://link.springer.com/10.1007/978-3-031-12282-8

Kaariainen, M. 2006. Semi-Supervised Model Selection Based on Cross-Validation. In: The 2006 IEEE International Joint Conference on Neural Network Proceedings. IEEE, Vancouver, BC, Canada. p. 1894–1899. Available from: http://ieeexplore.ieee.org/document/1716341/

Kavlak, A. T., M. Pastell, and P. Uimari. 2023. Disease detection in pigs based on feeding behaviour traits using machine learning. Biosyst. Eng. 226:132–143. doi:10.1016/j.biosystemseng.2023.01.004.

Kessner, D., and J. Novembre. 2015. Power Analysis of Artificial Selection Experiments Using Efficient Whole Genome Simulation of Quantitative Traits. 199:991–1005. doi:10.1534/genetics.115.175075/-/DC1.

Kichaev, G., et al. 2019. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am. J. Hum. Genet. 104:65–75. doi:10.1016/j.ajhg.2018.11.008.

Krahn, G. 2018. Predictors for wean-finish mortality events in a commercial swine production system [PhD Dissertation]. Iowa State University.

Kreitman, M. 2000. Methods to detect selection in populations with applications to the human. Annu. Rev. Genomics Hum. Genet. 1:539–559. doi:10.1146/annurev.genom.1.1.539.

Lagua, E., H.-S. Mun, K. M. B. Ampode, V. Chem, H.-R. Park, Y.-H. Kim, and C.-J. Yang. 2024. Monitoring using artificial intelligence reveals critical links between housing conditions and respiratory health in pigs. J. Anim. Behav. Biometeorol. 12:2024008. doi:10.31893/jabb.2024008.

Lasser, J., et al. 2021. Integrating diverse data sources to predict disease risk in dairy cattle—a machine learning approach. J. Anim. Sci. 99:skab294. doi:10.1093/jas/skab294.

Le Dividich, J., and P. Herpin. 1994. Effects of climatic conditions on the performance, metabolism and health status of weaned piglets: a review. Livest. Prod. Sci. 38:79–90. doi:10.1016/0301-6226(94)90052-3.

Li, Mingzhou, et al. 2017. Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. Genome Res. 27:865–874. doi:10.1101/gr.207456.116.

Li, Y., et al. 2023. A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering. Int. J. Appl. Earth Obs. Geoinformation. 118:103269. doi:10.1016/j.jag.2023.103269.

Losinger, W. C., E. J. Bush, M. A. Smith, and B. A. Corso. 1998. An analysis of mortality in the grower/finisher phase of swine production in the United States. Prev. Vet. Med. 33:121–145. doi:10.1016/S0167-5877(97)00052-4.

Ma, Y., J. Wei, Q. Zhang, L. Chen, J. Wang, J. Liu, and X. Ding. 2015. A Genome Scan for Selection Signatures in Pigs. PLOS ONE. 10:e0116850. doi:10.1371/journal.pone.0116850.

Macaulay, F. R. 1931. Introduction to "The Smoothing of Time Series." In: The Smoothing of Time Series. NBER. p. 17–31. Available from: http://www.nber.org/chapters/c9360

Maes, D. G. D., L. Duchateau, A. Larriestra, J. Deen, R. B. Morrison, and A. de Kruif. 2004. Risk Factors for Mortality in Grow-finishing Pigs in Belgium. J. Vet. Med. Ser. B. 51:321–326. doi:10.1111/j.1439-0450.2004.00780.x.

Magalhaes, E. S., et al. 2023. Field Implementation of Forecasting Models for Predicting Nursery Mortality in a Midwestern US Swine Production System. Animals. 13:2412. doi:https://doi.org/10.3390/ani13152412.

Magalhães, E. S., et al. 2024. Utilizing productivity and health breeding-to-market information along with disease diagnostic data to identify pig mortality risk factors in a U.S. swine production system. Front. Vet. Sci. 10:1301392. doi:10.3389/fvets.2023.1301392.

Maloof, M. A. 2003. Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown. In: Workshop on Learning from Imbalanced Data Sets II. Washington, DC.

Matthews, S. G., A. L. Miller, T. PlÖtz, and I. Kyriazakis. 2017. Automated tracking to measure behavioural changes in pigs for health and welfare monitoring. Sci. Rep. 7:17582. doi:10.1038/s41598-017-17451-6.

McVean, G. 2009. A Genealogical Interpretation of Principal Components Analysis. M. Przeworski, editor. PLoS Genet. 5:e1000686. doi:10.1371/journal.pgen.1000686.

Mehling, S., et al. 2019. Mortality Patterns in a Commercial Wean-To Finish Swine Production System. Vet. Sci. 6:49. doi:10.3390/vetsci6020049.

MetaFarms, and National Pork Board. 2022. Production Analysis Summary for U.S. Pork Industry: 2017-2021. Available from: https://porkcheckoff.org/research/production-analysis-summary-for-u-s-pork-industry-2017-2021/

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. 2023. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071). Available from: https://CRAN.R-project.org/package=e1071

Moon, S., et al. 2015. A genome-wide scan for signatures of directional selection in domesticated pigs. BMC Genomics. 16:130. doi:10.1186/s12864-015-1330-x.

Moreno-Torres, J. G., T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. 2012. A unifying view on dataset shift in classification. Pattern Recognit. 45:521–530. doi:10.1016/j.patcog.2011.06.019.

Morota, G., R. V. Ventura, F. F. Silva, M. Koyama, and S. C. Fernando. 2018. BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Machine learning and data mining advance predictive big data analysis in precision animal agriculture1. J. Anim. Sci. 96:1540–1550. doi:10.1093/jas/sky014.

Navada, A., A. N. Ansari, S. Patil, and B. A. Sonkamble. 2011. Overview of use of decision tree algorithms in machine learning. In: 2011 IEEE Control and System Graduate Research Colloquium. IEEE, Shah Alam, Malaysia. p. 37–42. Available from: http://ieeexplore.ieee.org/document/5991826/

Nti, I. K., O. Nyarko-Boateng, and J. Aning. 2021. Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation. Int. J. Inf. Technol. Comput. Sci. 13:61–71. doi:10.5815/ijitcs.2021.06.05.

Pennings, P. S., and J. Hermisson. 2006. Soft Sweeps II—Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. Mol. Biol. Evol. 23:1076–1084. doi:10.1093/molbev/msj117.

Perperoglou, A., W. Sauerbrei, M. Abrahamowicz, and M. Schmid. 2019. A review of spline function procedures in R. BMC Med. Res. Methodol. 19:46. doi:10.1186/s12874-019-0666-3.

Pessoa, J., et al. 2022. Environmental Risk Factors Influence the Frequency of Coughing and Sneezing Episodes in Finisher Pigs on a Farm Free of Respiratory Disease. Animals. 12:982. doi:10.3390/ani12080982.

Piñeiro, C., J. Morales, M. Rodríguez, M. Aparicio, E. G. Manzanilla, and Y. Koketsu. 2019. Big (pig) data and the internet of the swine things: a new paradigm in the industry. Anim. Front. 9:6–15. doi:10.1093/af/vfz002.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909. doi:10.1038/ng1847.

Purcell, S., et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81:559–575. doi:10.1086/519795.

Qanbari, S., and H. Simianer. 2014. Mapping signatures of positive selection in the genome of livestock. Livest. Sci. 166:133–143. doi:10.1016/j.livsci.2014.05.003.

R Core Team. 2023. R: A Language and Environment for Statistical Computing. Available from: https://www.R-project.org/

Rahman, M. T., T. M. Brown-Brandl, G. A. Rohrer, S. R. Sharma, V. Manthena, and Y. Shi. 2023. Statistical and machine learning approaches to describe factors affecting preweaning mortality of piglets. Transl. Anim. Sci. 7:txad117. doi:10.1093/tas/txad117.

Raj, A., M. Stephens, and J. K. Pritchard. 2014. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics. 197:573–589. doi:10.1534/genetics.114.164350.

Raudys, A., V. Lenčiauskas, and E. Malčius. 2013. Moving Averages for Financial Data Smoothing. Information and Software Technologies. Vol. 403. Springer Berlin Heidelberg, Berlin, Heidelberg. p. 34–45. Available from: http://link.springer.com/10.1007/978-3-642-41947-8_4

Rowan, T. N., H. J. Durbin, C. M. Seabury, R. D. Schnabel, and J. E. Decker. 2021. Powerful detection of polygenic selection and evidence of environmental adaptation in US beef cattle. PLoS Genet. 17:e1009652. doi:10.1371/journal.pgen.1009652.

Sagi, O., and L. Rokach. 2021. Approximating XGBoost with an interpretable decision tree. Inf. Sci. 572:522–542. doi:10.1016/j.ins.2021.05.055.

Saito, T., and M. Rehmsmeier. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. G. Brock, editor. PLOS ONE. 10:e0118432. doi:10.1371/journal.pone.0118432.

Silva, A. P. S. P., et al. 2022. Cough associated with the detection of Mycoplasma hyopneumoniae DNA in clinical and environmental specimens under controlled conditions. Porc. Health Manag. 8:6. doi:10.1186/s40813-022-00249-y.

Sofaer, H., J. Hoeting, and C. Jarnevich. 2018. The area under the precision-recall curve as a performancemetric for rare binary events. Methods Ecol. Evol. 10:565–577. doi:10.1111/2041-210X.13140.

Sollis, E., et al. 2023. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 51:D977–D985. doi:10.1093/nar/gkac1010.

Stas, N. M. 2017. Effect of sire line and selection index category on pig growth performance from weaning to harvest and carcass characteristics. University of Illinois at Urbana-Champaign.

Storey, J. D., A. J. Bass, A. Dabney, and D. Robinson. 2017. qvalue: Q-value estimation for false discovery rate control. Available from: http://github.com/StoreyLab/qvalue

Su, G., T. Liu, O. F. Christensen, M. S. Lund, and B. Nielsen. 2022. 778. Feasibility of reducing mortality of pigs from birth to slaughter by genetic selection. In: Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP). Wageningen Academic Publishers, Rotterdam, the Netherlands. p. 3204–3207. Available from: https://www.wageningenacademic.com/doi/10.3920/978-90-8686-940-4_778

Swets, J. A. 1988. Measuring the Accuracy of Diagnostic Systems. Science. 240:1285–1293. doi:10.1126/science.3287615.

Tay, J. K., B. Narasimhan, and T. Hastie. 2023. Elastic Net Regularization Paths for All Generalized Linear Models. J. Stat. Softw. 106. doi:10.18637/jss.v106.i01. Available from: https://www.jstatsoft.org/v106/i01/

Vabalas, A., E. Gowen, E. Poliakoff, and A. J. Casson. 2019. Machine learning algorithm validation with a limited sample size. E. Hernandez-Lemus, editor. PLOS ONE. 14:e0224365. doi:10.1371/journal.pone.0224365.

Van Klompenburg, T., and A. Kassahun. 2022. Data-driven decision making in pig farming: A review of the literature. Livest. Sci. 261:104961. doi:10.1016/j.livsci.2022.104961.

Varona, L., and D. Sorensen. 2010. A Genetic Analysis of Mortality in Pigs. Genetics. 184:277–284. doi:10.1534/genetics.109.110759.

Vázquez Diosdado, J. A., et al. 2015. Classification of behaviour in housed dairy cows using an accelerometer-based activity monitoring system. Anim. Biotelemetry. 3:15. doi:10.1186/s40317-015-0045-8.

Venables, W. N., and B. D. Ripley. 2002. Modern Applied Statistics with S. Fourth Edition. Springer, New York. Available from: https://www.stats.ox.ac.uk/pub/MASS4/

Vranken, E., and D. Berckmans. 2017. Precision livestock farming for pigs. Anim. Front. 7:32–37. doi:10.2527/af.2017.0106.

Walsh, B., M. Lynch, and M. Lynch. 2018. Evolution and selection of quantitative traits. Oxford University Press, New York, NY.

Wang, H., Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar. 2024. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. J. Big Data. 11:44. doi:10.1186/s40537-024-00905-w.

Warr, A., et al. 2020. An improved pig reference genome sequence to enable pig genetics and genomics research. GigaScience. 9:giaa051. doi:10.1093/gigascience/giaa051.

Wathes, C. M., H. H. Kristensen, J.-M. Aerts, and D. Berckmans. 2008. Is precision livestock farming an engineer's daydream or nightmare, an animal's friend or foe, and a farmer's panacea or pitfall? Comput. Electron. Agric. 64:2–10. doi:10.1016/j.compag.2008.05.005.

Weigand, H., and F. Leese. 2018. Detecting signatures of positive selection in non-model species using genomic data. Zool. J. Linn. Soc. 184:528–583. doi:10.1093/zoolinnean/zly007.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Available from: https://ggplot2.tidyverse.org

Wickham, H., et al. 2019. Welcome to the Tidyverse. J. Open Source Softw. 4:1686. doi:10.21105/joss.01686.

Wickham, H., R. Francois, L. Henry, K. Müller, and Vaughan. 2023. dplyr: A Grammar of Data Manipulation. Available from: https://CRAN.R-project.org/package=dplyr

Wilkinson, S., et al. 2013. Signatures of Diversifying Selection in European Pig Breeds. P. M. Visscher, editor. PLoS Genet. 9:e1003453. doi:10.1371/journal.pgen.1003453.

Wilson, J. R., and K. A. Lorenz. 2015. Short History of the Logistic Regression Model. In: Modeling Binary Correlated Responses using SAS, SPSS and R. Springer International Publishing, Cham. p. 17–23. Available from: https://doi.org/10.1007/978-3-319-23805-0_2

Yang, J., B. et al. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42:565–569. doi:10.1038/ng.608.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am. J. Hum. Genet. 88:76–82. doi:10.1016/j.ajhg.2010.11.011.

Yang, S., X. Li, K. Li, B. Fan, and Z. Tang. 2014. A genome-wide scan for signatures of selection in Chinese indigenous and commercial pig breeds. BMC Genet. 15:7. doi:10.1186/1471-2156-15-7.

Zhi-Hua Zhou and Xu-Ying Liu. 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans. Knowl. Data Eng. 18:63–77. doi:10.1109/TKDE.2006.17.

Zou, H., and T. Hastie. 2005. Regularization and Variable Selection Via the Elastic Net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67:301–320. doi:10.1111/j.1467-9868.2005.00503.x.

VITA

Caleb Grohmann is the Production Data Scientist at Carthage Veterinary Service, LTD. Caleb completed a Ph.D. with an emphasis in Bioinformatics at the University of Missouri Institute for Data Science and Informatics. Grohmann received his B.S. at the University of Missouri – Columbia, where he majored in Animal Sciences with a minor in Agricultural Economics. He then earned his M.S. in Animal Sciences at the University of Illinois at Urbana – Champaign, where his research focused on estimating relationships between growth performance, carcass, and meat quality measurements and carcass and primal cut value of weaning-to-finishing pigs. His Ph.D. research focused on the prediction of mortality episode occurrence in wean-to-finish pig farms using machine learning models.