Survival Analysis of Canines with Crushing's syndrome

- Shrinidhi Rajesh

Introduction

Hyperadrenocorticism, commonly known as Cushing's syndrome, is a complex endocrine disorder characterized by excessive cortisol production in dogs. With an estimated prevalence of 0.28% in primary care practice, it represents a significant health concern in the canine population. The disease primarily affects older dogs, with an average age at diagnosis of nine years in primary care settings. Previous studies on canine hyperadrenocorticism have largely focused on referral populations, potentially limiting the generalizability of findings to broader primary care settings. Median survival times for dogs with pituitary-dependent hyperadrenocorticism (PDH) treated with trilostane in referral settings have been reported to range from 662 to 900 days, while for adrenal-dependent hyperadrenocorticism (ADH), survival times range between 353 and 475 days. However, these figures may not accurately reflect outcomes in primary care practice.

My study aims to address this knowledge gap by analyzing a dataset of 219 dogs diagnosed with hyperadrenocorticism across multiple primary care practices in England. This research will provide valuable insights into the natural history of the disease and identify key prognostic factors in a real-world clinical setting. Trilostane (Vetoryl Capsules, Dechra Veterinary Products) is the primary pharmacological intervention for canine hyperadrenocorticism in the UK. However, the impact of trilostane treatment and dose adjustments on survival in primary care settings remains to be fully elucidated. Additionally, the influence of comorbidities such as diabetes mellitus, urinary tract infections, and other concurrent conditions on survival outcomes warrants further investigation.

To comprehensively analyze the survival characteristics and prognostic factors, this study will employ a range of statistical techniques including Kaplan-Meier survival analysis, log-rank tests, Cox proportional hazards regression, and random survival forests. These methods will allow for a nuanced examination of survival times, treatment effects, and the impact of various clinical and demographic factors on disease outcomes. By focusing on a primary care population, this study aims to provide clinically relevant insights that can inform evidence-based decision-making in veterinary practice. The findings will contribute to a better understanding of canine hyperadrenocorticism management in real-world settings, potentially improving treatment strategies and patient outcomes for affected dogs.

Dataset Description

The study will utilize a dataset of 219 canine patients diagnosed with hyperadrenocorticism (HAC), collected from multiple primary care veterinary practices across England. The dataset includes key patient demographics, diagnostic information, treatment details, disease type, comorbidities, and survival outcomes. These variables allow for an in-depth analysis of survival trends, treatment effectiveness, and the factors influencing the prognosis of dogs with HAC in a real-world primary care setting.

Attribute	Description
PatientID	**Unique identifier for each dog**
Site	**Veterinary practice ID (110 primary care practices in England)**
BirthDate	**Dog's date of birth**
Sex	**Male / Female**
IsNeutered	**Yes (neutered) / No (intact)**
Weight	**Maximum recorded body weight in kg**
BreedRelativeWeight	**1, 2, 3, or 4 (comparison to breed average)**
Breed	**Specific breed, Breed Unspecified or Crossbreed**
KC_Group	**Kennel Club breed group**

Purebreed_Status	**1 (Purebred) / 2 (Crossbreed/ Unspecified)**
Insurance	**Insured / Uninsured / Unknown (0)**
Date of First Suspicion_4	**Date HAC first suspected**
Date of Diagnosis_3	**Date of HAC diagnosis**
Pre-ACTH	**Cortisol level (nmol/l) before ACTH stimulation test**
Post-ACTH	**Cortisol level (nmol/l) after ACTH stimulation test**
Date Trilostane Started_5	**Start date of trilostane treatment**
Treated with Trilostane_7	**1 (Yes) / 2 (No)**
Trilostane Starting Dose (mg/kg)_8	**Initial trilostane dose**
Trilostane SID/BID_9	**SID (once daily) / BID (twice daily) / Unknown / NA**
Changes to Trilostane_6	**1, 2, 3, or 4 (dose adjustments)**
Stay_vs_Stop	**1 (Stay on treatment) / 2 (Stop treatment)**
Died	**1 (Died during study) / 0 (Alive or lost to follow-up)**
How Died	**Euthanasia / Not known / Unknown / Unassisted**
FailureDate	**Date of death or last clinical visit**
Censored_10	**Yes (alive/lost to follow-up) / No (died)/ NA**
Why Censored_11	**Moved practice / No record in last 3 months / Alive / Practice cannot contact / NA**
Neuro Signs	**Yes / No**
Complications	**Yes / Unknown**
Hypertensive_Yes4	**Yes / No / Unknown**
Oversuppression?	**Yes / No / Unknown**
Cortisol Stayed <250	**Yes / No / Unknown**
Cortisol Went <40	**Yes / No / Unknown**
Number_Comorbidities	**0, 1, 2, 3, 4, or 5**
comorb_UTI	**Yes / No**
comorb_dm	**Yes / No**
comorb_hypot	**Yes / No**

Research Objectives

This study focuses on exploring key aspects of survival in dogs diagnosed with hyperadrenocorticism within primary care settings. The specific objectives are as follows:

- Median Survival Time: Determine the median survival time for dogs diagnosed with hyperadrenocorticism in primary care settings.
- Effect of Trilostane Treatment: Evaluate the impact of trilostane treatment on survival by comparing treated and untreated dogs.
- Purebred Status and Neutering: Assess the effects of purebred status, breeds and neutering on survival outcomes.
- Prognostic Factors: Identify important prognostic factors influencing survival, including age, weight, breed, and disease type (pituitary-dependent vs. adrenal-dependent).
- Comorbidity Impact: Analyze the influence of comorbidities such as diabetes, hypertension, and urinary tract infections on survival outcomes.

By addressing these objectives, the study aims to provide valuable insights into the survival characteristics and key determinants of outcomes in canine hyperadrenocorticism, improving evidence-based management in primary care practice.

Statistical Methods Utilized

To analyze survival times and identify key prognostic factors in dogs diagnosed with hyperadrenocorticism, the following statistical techniques will be employed:

Kaplan-Meier Survival Analysis:

Kaplan-Meier survival analysis will be utilized to estimate the overall survival curve for dogs diagnosed with hyperadrenocorticism in primary care settings. This technique will allow for the visualization of survival probabilities over time and provide an understanding of the disease's natural history. The method will also be applied to compare survival between different groups, such as treated versus untreated dogs and purebred versus non-purebred dogs, to assess differences in survival outcomes.

Log-Rank Test:

The log-rank test will be used to compare the survival distributions between various groups, including those receiving trilostane treatment and those that do not, as well as groups based on breed, neutering status, and comorbidity presence (e.g., diabetes, hypertension). This test will evaluate whether differences in survival rates between these groups are statistically significant, providing insights into factors that may influence survival time in primary care settings.

Cox Proportional Hazards Regression:

Cox proportional hazards regression will be employed to identify and quantify significant prognostic factors that influence survival, such as age, weight, breed, disease type (pituitary-dependent vs. adrenal-dependent), and the presence of comorbidities. By modeling survival as a function of these covariates, this technique will assess how each factor affects the hazard rate (the risk of death) over time, enabling a deeper understanding of the variables that significantly impact survival outcomes.

Random Survival Forests:

Random survival forests will be applied to explore non-linear relationships and interactions between various prognostic factors, offering an alternative to Cox proportional hazards regression. This method will allow for the identification of the most influential variables on survival outcomes without assuming linearity or proportional hazards. It will provide an enriched understanding of the factors contributing to survival and reveal complex patterns that may not be captured by traditional regression models.

By using these statistical methods, the study aims to comprehensively analyze survival trends and identify the key factors influencing survival outcomes in dogs with hyperadrenocorticism, ultimately contributing to more informed decision-making in primary care veterinary practice.

Analysis and Preprocessing

Preprocessing is a critical step in any data analysis project, as it ensures that the dataset is clean, consistent, and properly formatted for modeling. In the context of survival analysis, preprocessing addresses issues such as missing values, variable transformations, and the correct encoding of categorical variables. This step is essential for ensuring that the data is compatible with the models used, such as the Kaplan-Meier estimator, Cox Proportional Hazards (CoxPH) model, and

Random Survival Forests (RSF), and that the analysis produces reliable and accurate results. By handling these tasks, preprocessing helps prevent biases or inconsistencies that could distort the analysis.

Although preprocessing can be time-consuming, it is crucial for ensuring high-quality analysis. The process typically involves cleaning the data, transforming variables to fit model requirements, and addressing time-related aspects that are unique to survival analysis. The length of this step depends on the complexity and quality of the dataset, but it is a necessary investment to improve model accuracy and ensure the validity of the findings. Ultimately, well-executed preprocessing is key to the success of survival analysis, allowing the models to function effectively and produce trustworthy insights.

Preprocessing Approach:

The preprocessing of the data will be carried out in two distinct stages.

The first stage involves an initial data analysis and cleaning, where the dataset will be examined for unnecessary columns, missing values (nulls), and other inconsistencies. Any irrelevant or redundant columns will be removed, and appropriate strategies will be applied to handle missing data, ensuring that the dataset is cleaned and ready for further analysis.

The second stage focuses on preparing the data specifically for model fitting. This will include transforming categorical variables into suitable encodings (e.g., one-hot encoding or label encoding) to ensure that the data is in an appropriate format for fitting the Cox Proportional Hazards (CoxPH) and Random Survival Forests (RFS) models. This step is essential for converting non-numeric variables into numeric forms, enabling the models to properly handle and interpret the data.

Preprocessing – Part 1

The initial preprocessing step involved analyzing the dataset, which originally contained 38 columns, and removing unnecessary or irrelevant features. After carefully inspecting the columns, the last column was found to be extraneous and was removed, reducing the total number of columns to 37. Following this, a subset of the dataset was created by excluding the 'PatientID' column, which is not needed for survival analysis. The dataset now consists of 219 rows and 36 columns, with each row representing a unique canine patient diagnosed with hyperadrenocorticism. This subset will serve as the foundation for subsequent survival analyses, such as Kaplan-Meier and Log-Rank tests.

The next step involved handling missing values across the dataset. After inspecting the columns for null values, it was found that several columns contained missing data, with the most notable being 'Date trilostane started', 'Changes to trilostane', and 'Trilostane starting dose', which had 14, 13, and 20 missing values, respectively. Other columns such as 'Cause of death' and 'Censored' also had significant missing data, with 42 and 174 missing entries, respectively. Given the importance of these variables for survival analysis, appropriate strategies were considered to handle the missing values, including potential imputation or removal of columns depending on the proportion of missing data. The dataset now consists of 219 rows and is ready for the next steps, where we ensure no further issues arise due to missing data.

Then, the dataset underwent a thorough review to ensure it was prepared for modeling. This included converting the 'Died' column from a float to an integer type, ensuring that the data was compatible with the models to be used in survival analysis. The column now correctly reflects binary values (1 for death and 0 for survival). Additionally, variables such as 'site', which did not provide significant information for the analysis, were dropped to streamline the dataset. Following these steps, the data was checked for consistency, ensuring that the number of rows (219) remained unchanged, and the column names were appropriately aligned.

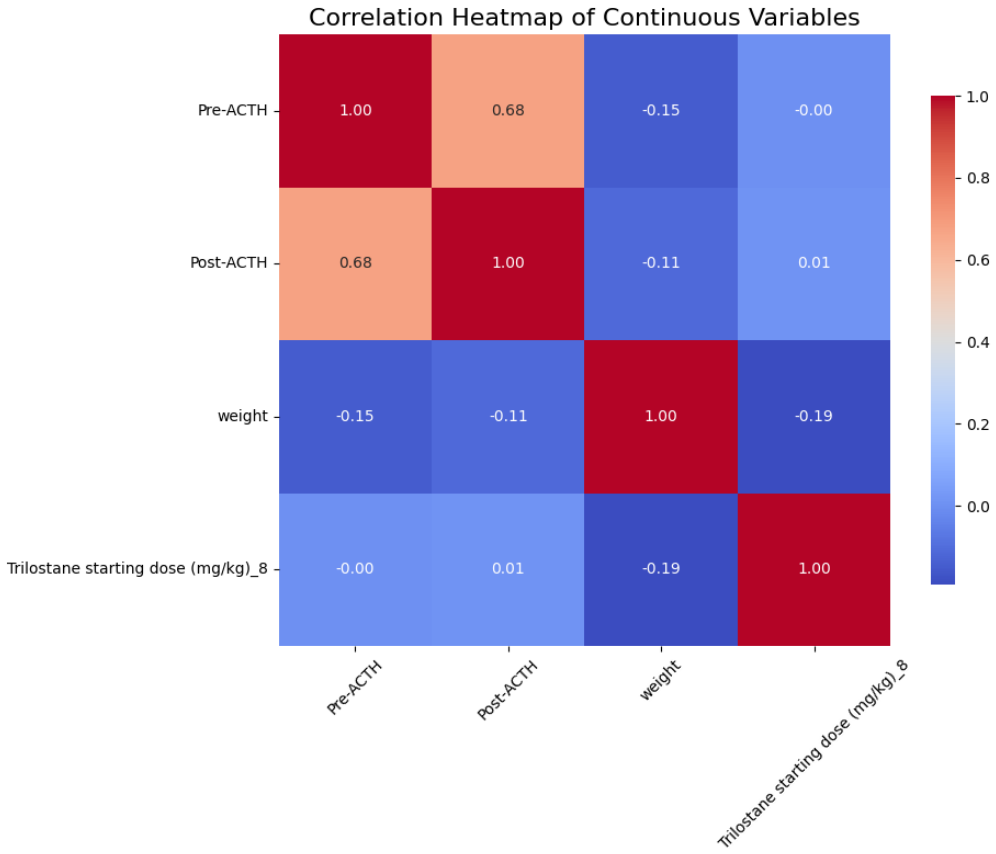
The dataset contains both categorical and continuous variables, each requiring specific handling to uncover insights. For the categorical variables, I converted the relevant columns to

categorical data types and calculated descriptive statistics, which revealed the number of unique entries for each category. Notable observations from the categorical data include a disproportionate number of entries for the "Died" column, where a large portion of the animals had the outcome labeled as "Died," with many categorized as "Euthanized." The "Breed" category showed high diversity, with "Crossbreed" and several terrier breeds being the most frequent. The "Cause of death" column had many entries marked as "Unknown," indicating missing or incomplete data that needs further examination or clarification.

Yes/No	Died
1	179
0	40

For the continuous variables, I calculated descriptive statistics, such as mean, standard deviation, minimum, and maximum values, for variables like "Pre-ACTH," "Post-ACTH," "weight," and "Trilostane starting dose." These descriptive statistics provided a better understanding of the distributions of these numerical values. Additionally, a correlation heatmap was generated to visualize the relationships between the continuous variables, highlighting how certain variables are correlated with each other. For example, it was noted that variables like "Pre-ACTH" and "Post-ACTH" are closely related, suggesting a potential interaction between these factors that could be significant for further analysis.

Statistics	Pre-ACTH	Post-ACTH	weight	Trilostane starting dose (mg/kg)_8
count	219	219	219	199
mean	53.794064	301.045662	16.314064	3.301508
std	101.19117	440.164489	10.872733	1.275695
min	0	0	0	1
25%	0	0	9.2	2.5
50%	0	0	12.75	3
75%	88.85	682.5	20.35	4
max	508	1380	65.6	8



The dataset also contained date-related columns that required conversion into a proper datetime format for analysis. Once converted, I created new time-related columns to calculate the periods between key events like birth, suspicion, diagnosis, treatment, and failure. This allowed for a more detailed view of the time intervals between important milestones in the clinical process for each animal. These new time-based features are valuable for understanding the timeline and progression of the conditions being studied.

Furthermore, I filtered the data to focus on uncensored cases, where "Died" is marked as 1, which means the animals had the "Died" outcome. This subset was created to examine the characteristics and factors that contributed to death, providing insights into the variables that may be important in predicting the outcome.

At the end of the preprocessing stages, the dataset is now in a clean and structured format, ready for modeling. In Part 1, irrelevant columns were removed, missing values were addressed, and categorical variables were converted into appropriate formats, ensuring compatibility with the survival analysis models. The handling of missing data, including strategies like imputation or removal, ensured that no critical variables were compromised, which is vital for survival analysis models like Kaplan-Meier and Cox Proportional Hazards (CoxPH).

In Part 2, transformations were applied to key variables, such as age at diagnosis and time from suspicion to diagnosis, and categorical variables were encoded for numeric compatibility. The final datasets, including the final_cox_data, canine_comorb, and random_data, were created to reflect the necessary clinical, treatment, and demographic information required by both the CoxPH and Random Survival Forest models.

Through these preprocessing efforts, the dataset is prepared to be used for advanced survival analysis, enabling reliable model fitting and the extraction of meaningful insights regarding the survival outcomes of dogs diagnosed with hyperadrenocorticism. The next steps will focus on

fitting the models and performing survival analysis to investigate the factors that influence patient outcomes.

Preprocessing - Part 2

The preprocessing steps for fitting the Cox Proportional Hazards (CoxPH) model and Random Survival Forests (RSF) involve creating new variables based on existing columns. The first transformation calculates the age at diagnosis by subtracting the birthdate from the diagnosis date and converting it to years. Similarly, the time from suspicion to diagnosis is calculated by subtracting the date of first suspicion from the diagnosis date. Both new variables are then rounded to two decimal places for improved readability. The resulting data is then verified to ensure the transformations are accurate.

In addition to these transformations, categorical and continuous columns are summarized to ensure they are properly prepared for modeling. Categorical columns, such as 'Died', 'Neuro signs', and 'Sex', are analyzed to check for the distribution of unique values, counts, and percentages. Null values are also reported to identify missing data. For continuous columns like 'Age_at_diagnosis(Years)' and 'Time_suspicion_to_diagnosis(Days)', summary statistics such as the mean, median, min, max, and standard deviation are calculated. These summaries help in understanding the distribution and potential issues within the data.

The summarized statistics and distributions of the variables are then displayed to assess the data's quality and completeness. For categorical variables, the unique values and their frequencies are shown along with the percentage of total entries they represent. Continuous variables are reviewed for outliers and central tendencies, ensuring that the data is ready for further modeling. These preprocessing steps are essential for the proper fitting of both CoxPH and RSF models, which rely on accurate and well-prepared data for survival analysis.

The dataset for the Cox Proportional Hazards (CoxPH) model was finalized by selecting relevant columns from the original canine dataset. This included clinical factors such as 'Pre-ACTH', 'Post-ACTH', and 'Neuro signs', along with demographic factors like 'Sex' and 'BreedRelativeWeight'. Treatment-related factors, such as whether the dog was treated with trilostane, and time-related variables, including 'Age_at_diagnosis(Years)' and 'Time_suspicion_to_diagnosis(Days)', were also included. The target outcome columns, 'Died' and 'Days_Diagnosis_to_Failure', were part of the final dataset, which was subsequently cleaned and prepared for the CoxPH model.

To ensure the data was ready for analysis, categorical variables were encoded into numeric values. For instance, the 'Sex' column was transformed by assigning 1 for 'Female' and 0 for 'Male', while other columns like 'Neuro signs' and 'Hypertensive_Yes4' were encoded as binary values (0 and 1). Columns such as 'Complications' with 'Unknown' values were handled by assigning a specific code. After encoding, columns were converted to the appropriate data types, and missing values were addressed, ensuring that the dataset was consistent and complete. The resulting final_cox_data dataset, containing 219 rows and relevant predictor variables, was now ready for input into the Cox Proportional Hazards model.

A new DataFrame, canine_comorb, was then created which focuses on comorbidity-related data. The three categorical variables related to comorbidities (comorb_UTI, comorb_dm, comorb_hypot) are converted into binary format, where 'Yes' is mapped to 1 and 'No' is mapped to 0. This transformation is done because machine learning models typically require numerical inputs, and binary encoding of categorical data allows the model to handle these variables effectively. Additionally, other relevant numerical columns such as Number_comorbidities, Days_Diagnosis_to_Failure, and Died are directly included in the new DataFrame. These steps help in preparing the data for further analysis, where the binary comorbidity variables and outcome data will be used to assess relationships between comorbidities and survival outcomes.

Following this, a new DataFrame random_data is created, which includes a more comprehensive set of variables for the Random Survival Forest model. This DataFrame combines both clinical data and treatment information, such as Neuro signs, Complications, Hypertensive_Yes4, as well as comorbidity variables like comorb_UTI, comorb_dm, and comorb_hypot. It also incorporates demographic and clinical treatment variables like Sex, Isneutered, and hormone levels (Pre-ACTH, Post-ACTH). The inclusion of these features allows the model to account for a wide range of factors that may influence the survival time or the likelihood of an event (in this case, death or failure). This ensures that the Random Survival Forest model can consider all relevant data for predicting survival outcomes and handling right-censored data effectively.

Methodolody and Results:

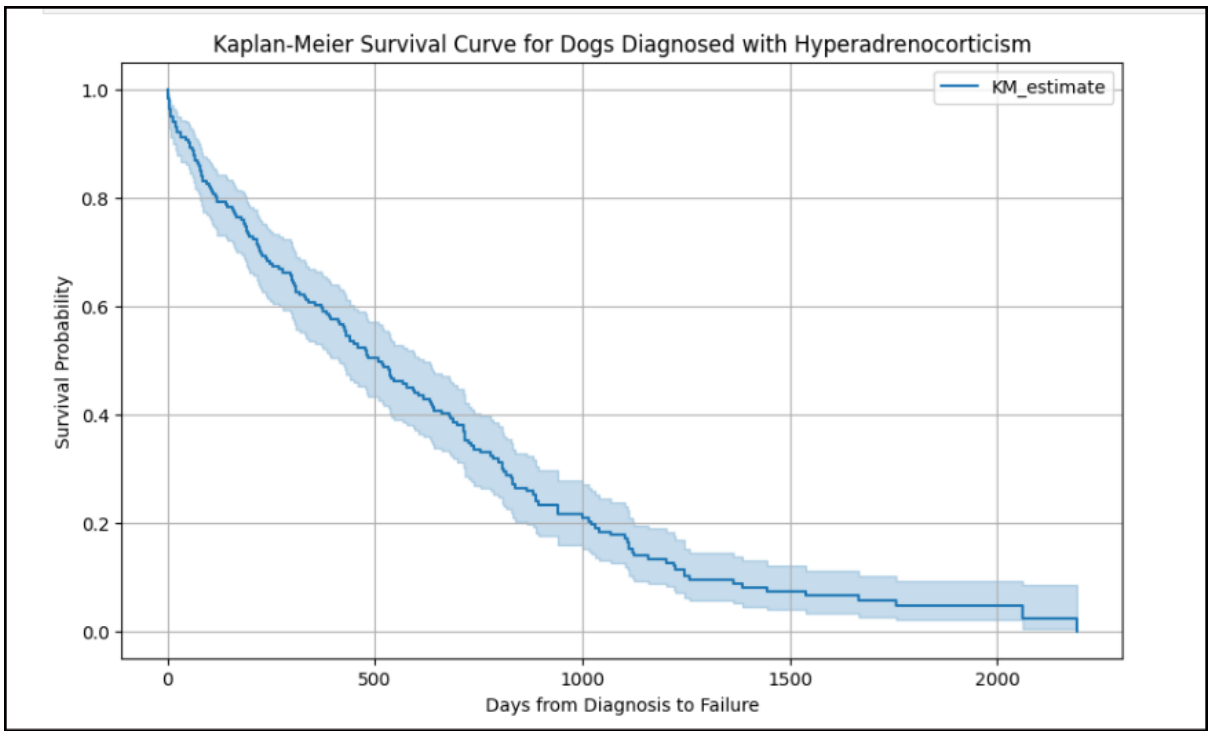
Kaplan-Meier Estimator Fitting

The first step in the analysis involves fitting the Kaplan-Meier estimator using the KaplanMeierFitter from the lifelines library. This estimator is used to estimate the survival function from the data, which describes the probability of survival over time for a population. In this case, the analysis is applied to a dataset of dogs diagnosed with hyperadrenocorticism. The duration variable is Days_Diagnosis_to_Failure, which represents the number of days from diagnosis to failure (or death), while the event indicator is the Died column, marking whether the dog has died (1) or survived (0). This fitting helps understand the overall survival trends and time-to-event characteristics of the group.

After fitting the Kaplan-Meier estimator, the survival function is plotted to visualize how the probability of survival changes over time. The x-axis represents the time (in days), while the y-axis shows the survival probability, which starts at 1 (100% survival) and decreases over time as more events (deaths) occur. This plot is essential for visually assessing the survival distribution of the cohort, indicating how the likelihood of survival drops as time progresses. It provides an immediate, visual understanding of the survival dynamics of dogs with hyperadrenocorticism.

The next step calculates the number of expected survivors at specific survival probabilities ranging from 1.0 (100% survival) to 0.0 (no survivors). For this, the survival function is examined at different thresholds (e.g., 0.9, 0.8, etc.), and the corresponding time at which the survival probability is closest to these thresholds is identified. This analysis helps translate survival probabilities into actionable insights. For instance, at a survival probability of 0.5 (median survival), around 110 dogs are expected to survive. These results help estimate the survival expectations for dogs at various probabilities and provide more granular insights into the prognosis for different time frames after diagnosis.

The output of the analysis is summarized in a table showing the survival probabilities, corresponding days, and expected survivors at each probability level. For example, at a 90% survival probability, the expected survival time is approximately 54 days, and 197 dogs are expected to survive up to that point. At a 50% survival probability, the median survival time is 510 days, with 110 dogs expected to survive beyond that point. This suggests that a significant number of dogs live past the median time, and survival rates decrease gradually over time. The median survival time is particularly important as it represents the time at which half of the dogs are expected to survive and half are expected to fail. In this case, the median survival time of 510 days gives a key benchmark for understanding the prognosis for dogs diagnosed with hyperadrenocorticism.

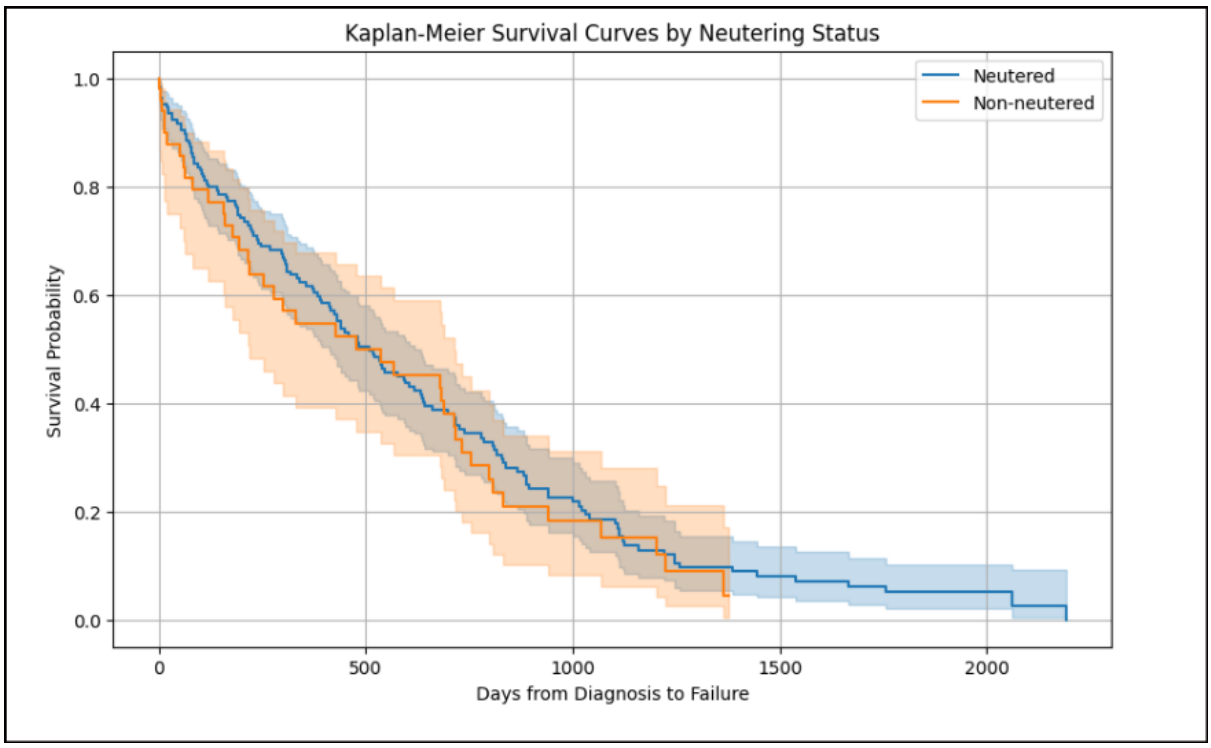


Log-Rank Test and Hypothesis Testing for Netuering

The analysis begins with performing a Log-Rank test to compare the survival distributions of neutered and non-neutered dogs. The Log-Rank test is a statistical test used to determine if there is a significant difference in the survival distributions of two or more groups. In this case, the test compares the survival times of dogs based on their neutering status (Isneutered). The p-value of the test is calculated, and since the p-value is 0.4999, which is greater than the commonly used significance level of 0.05, the result suggests there is no statistically significant difference in survival between neutered and non-neutered dogs.

A Bonferroni correction is applied to account for multiple comparisons, though here only one comparison is made between two groups. The corrected significance threshold is calculated as 0.05, which confirms that the p-value of 0.4999 is not significant. The next step calculates the survival probabilities for both groups at various thresholds (from 1.0 to 0.0). The survival probabilities are plotted for both neutered and non-neutered dogs using Kaplan-Meier survival curves. This step helps visualize the survival experience for each group over time and facilitates comparison.

The expected survival probabilities for neutered and non-neutered dogs are presented in tables. For neutered dogs, the survival probabilities decline gradually with time, with the majority of dogs surviving up to the median survival time of approximately 510 days. In contrast, non-neutered dogs show a faster decline in survival probabilities, with the majority expected to survive up to a median of 536 days. The results indicate that, although survival differs between the groups, the Log-Rank test suggests this difference is not statistically significant, implying that neutering does not have a strong impact on survival outcomes for dogs diagnosed with hyperadrenocorticism.

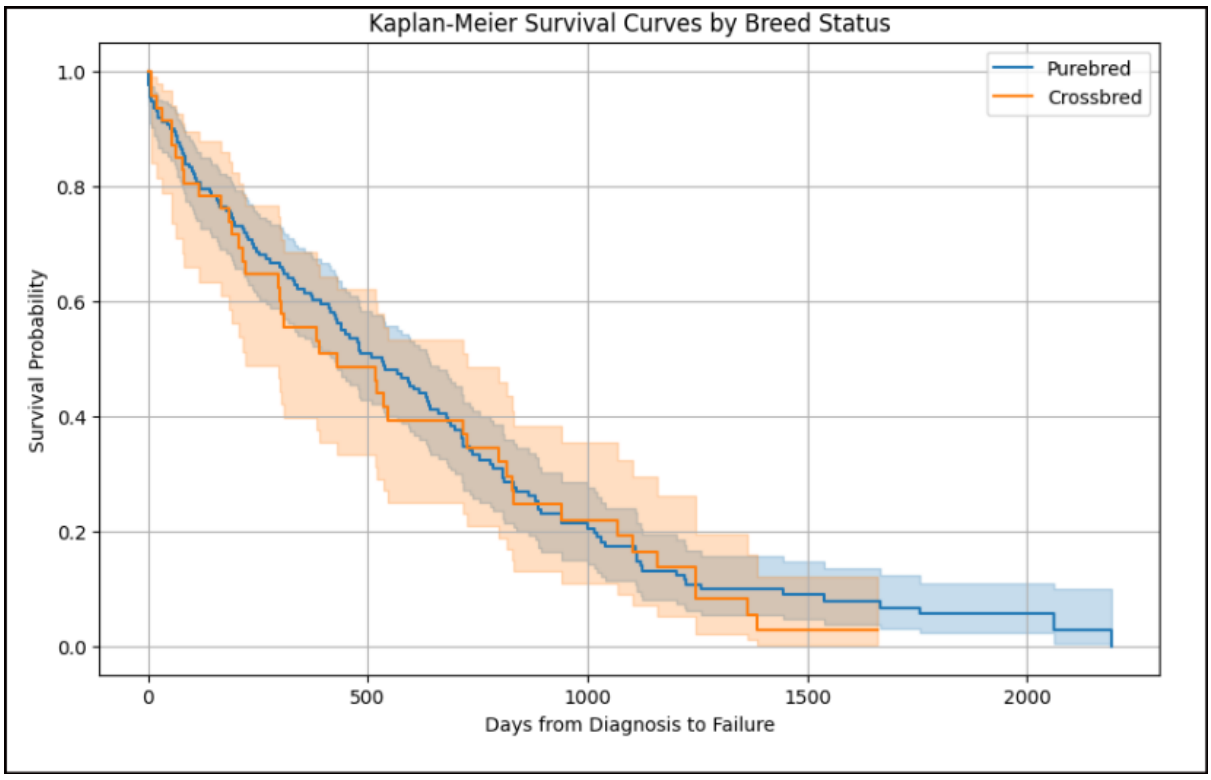


Log-Rank Test and Hypothesis Testing for Purebred Status

A Log-Rank test was conducted to compare the survival distributions of purebred and crossbred dogs based on their survival times from diagnosis to failure. The test was performed to evaluate if breed status influences survival outcomes. The resulting p-value of 0.6397 suggests that there is no significant difference in survival times between purebred and crossbred dogs. This indicates that breed status, as classified here, does not have a strong impact on survival after diagnosis, based on the data from this analysis.

The Bonferroni correction, though not needed here given only one comparison, was calculated to confirm that the significance threshold for this test remains 0.05. After performing the Kaplan-Meier survival analysis, survival probabilities were calculated for both purebred and crossbred groups at various survival thresholds ranging from 1.0 to 0.0. The survival curves for both groups were plotted, showing the proportion of survivors over time. The curves indicate that while survival trends for both groups decline over time, the actual survival probabilities are quite similar, further supporting the result of the Log-Rank test.

The expected survival probabilities were calculated for both purebred and crossbred dogs, providing a detailed breakdown of survival over time for different survival thresholds. For purebred dogs, the survival probabilities decrease gradually, with 172 dogs initially surviving at the highest probability, compared to 47 crossbred dogs. However, the survival times and number of expected survivors at each threshold are not drastically different between the two groups, supporting the conclusion from the Log-Rank test that breed status does not significantly impact survival. This suggests that factors other than breed might be more influential in determining survival outcomes for these dogs.

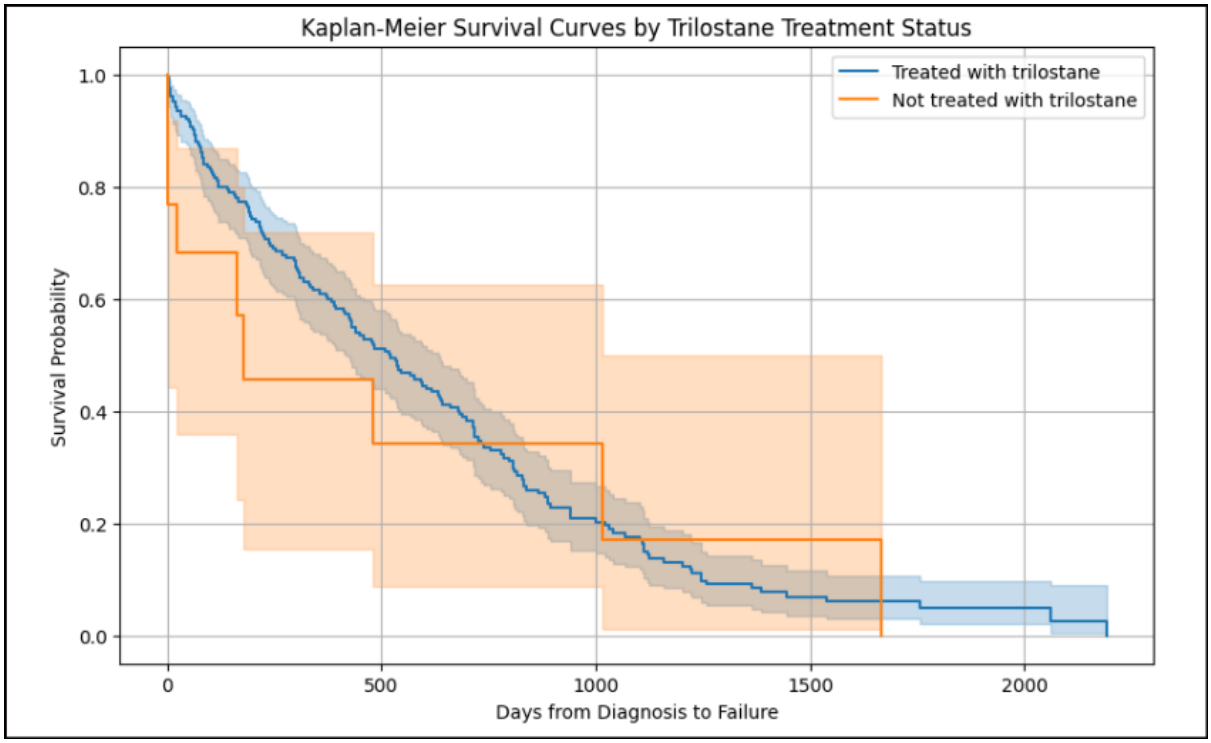


Log-Rank Test and Hypothesis Testing for Trilostane Treatment Status

A Log-Rank test was conducted to compare the survival distributions of dogs treated with trilostane and those not treated with trilostane, based on their survival times from diagnosis to failure. The test was performed to assess if trilostane treatment influences survival outcomes. The resulting p-value of 0.5896 suggests that there is no significant difference in survival times between the treated and not treated groups. This indicates that trilostane treatment, as classified in this analysis, does not have a strong impact on survival after diagnosis, based on the data.

The Bonferroni correction, although not required in this case given only one comparison, was calculated to confirm that the significance threshold for this test remains 0.05. After performing the Kaplan-Meier survival analysis, survival probabilities were calculated for both treated and non-treated groups at various survival thresholds ranging from 1.0 to 0.0. The survival curves for both groups were plotted, showing the proportion of survivors over time. These curves demonstrate that, while both groups experience a decline in survival over time, the survival probabilities for treated and non-treated dogs are similar, further supporting the conclusion from the Log-Rank test.

The expected survival probabilities were also calculated for both groups, providing a breakdown of survival over time for different survival thresholds. For the treated group, survival probabilities decrease gradually, with 206 dogs initially surviving at the highest probability, compared to only 13 dogs in the non-treated group. However, the survival times and the number of expected survivors at each threshold are not drastically different between the two groups, reinforcing the result from the Log-Rank test that trilostane treatment does not significantly affect survival. This suggests that factors other than trilostane treatment might be more influential in determining survival outcomes for these dogs.



Log-Rank Test for Pairwise Breed Comparisons

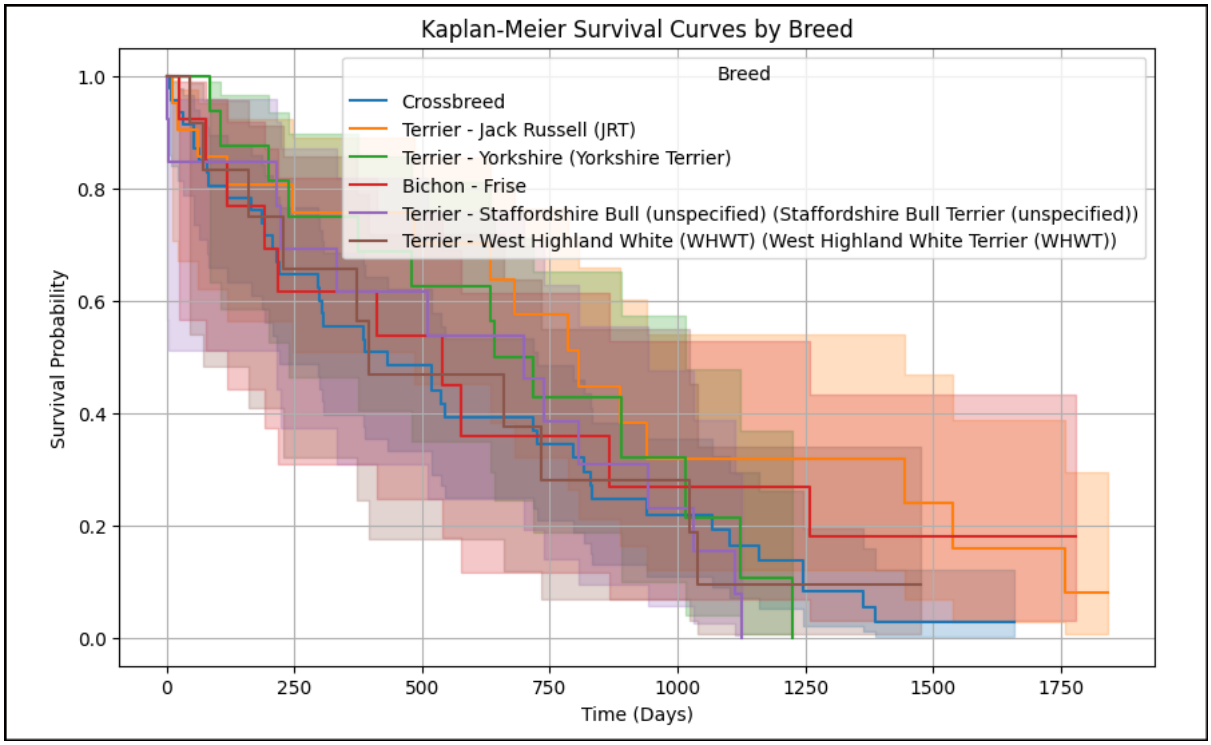
A Kaplan-Meier survival analysis was conducted to evaluate the survival distributions of different dog breeds with sufficient representation (at least 10 samples). The survival curves for each breed were estimated, showing the survival probability over time. The curves provide insights into the likelihood of survival across various time points, with each breed exhibiting unique survival patterns. The survival probabilities generally decrease over time for all breeds, but some breeds exhibit different rates of survival, indicating potential differences in longevity or resilience.

The Log-Rank test was performed to compare survival times between pairs of breeds. For the comparison between Crossbreed and Terrier - Jack Russell (JRT), the p-value was 0.0364, which is statistically significant ($p < 0.05$), suggesting a notable difference in survival between these two breeds.

In contrast, the comparison between Crossbreed and Terrier - Yorkshire (Yorkshire Terrier) yielded a p-value of 0.6692, which is not significant ($p > 0.05$), indicating that survival times between these breeds are similar. Similarly, comparisons between Crossbreed and Bichon - Frise (p-value = 0.3684), Crossbreed and Terrier - Staffordshire Bull (unspecified) (p-value = 0.7710), and Crossbreed and Terrier - West Highland White (WHWT) (p-value = 0.8338) also showed no significant differences, with p-values greater than the threshold of 0.05.

The analysis of Terrier - Jack Russell (JRT) versus Terrier - Yorkshire (Yorkshire Terrier) resulted in a p-value of 0.2577, indicating no significant difference in survival between these two breeds. Likewise, the comparison between Terrier - Jack Russell (JRT) and Bichon - Frise (p-value = 0.5159), Terrier - Jack Russell (JRT) and Terrier - Staffordshire Bull (unspecified) (p-value = 0.1214), and Terrier - Jack Russell (JRT) and Terrier - West Highland White (WHWT) (p-value = 0.2466) also showed no significant survival differences.

Overall, the Log-Rank tests suggest that while there are some breed pairs with significant survival differences (such as between Crossbreed and Terrier - Jack Russell (JRT)), many comparisons reveal no substantial differences in survival, as evidenced by the high p-values across several breed pairs. These findings highlight that breed differences in survival times may be more nuanced and not uniformly significant across all pairs.



Cox Proportional Hazard Model

In this analysis, a Cox Proportional Hazards (CoxPH) model was applied to examine factors influencing the survival times of dogs based on various clinical and demographic variables. The dataset, final_cox_data, included both continuous and categorical features such as age, weight, number of comorbidities, and specific medical conditions. Before fitting the model, it was essential to ensure that categorical variables were encoded into numerical values, as the CoxPH model requires all input variables to be numeric. This encoding process involved transforming categorical variables like 'Sex', 'BreedRelativeWeight', and 'Purebreed_status' into numerical representations so that they could be appropriately included in the model.

The CoxPH model was then fitted using the lifelines package, with 'Days_Diagnosis_to_Failure' as the duration column and 'Died' as the event column. A summary of the model was generated, providing coefficients, hazard ratios (exp(coef)), and statistical significance for each variable. Notably, some variables showed significant effects on survival times. For instance, 'Neuro signs' and 'Age_at_diagnosis(Years)' had significant hazard ratios, suggesting that neurological signs and age at diagnosis are important predictors of survival. On the other hand, 'Cortisol stayed <250' and 'Number_comorbidities' showed less significance, indicating their weaker relationship with survival in this model.

The model also underwent a test for proportional hazards assumptions using cph.check_assumptions(), which revealed no violations. This means that the model correctly accounts for the assumption that the hazard ratios are constant over time, a key condition for CoxPH models. As a result, the interpretation of the model coefficients as hazard ratios is valid. The concordance index (C-index) was 0.66, which indicates a moderate ability of the model to predict survival times based on the input variables.

Finally, partial hazards were predicted for each individual using the fitted CoxPH model. Scatter plots were generated to visualize the relationship between each continuous variable (such as age, weight, and time from suspicion to diagnosis) and the predicted partial hazard.

CoxPH Model on Comorbidities

The model was applied to this dataset to examine how comorbidities, like UTI, diabetes, and hypotension, along with the number of comorbidities, impact the survival time (Days_Diagnosis_to_Failure) and whether the dog dies (Died) during the study period.

The results of the model indicate that having a UTI significantly increases the risk of failure, with a hazard ratio of 1.82, meaning that dogs with a UTI are 82% more likely to experience failure (death) compared to those without it. The presence of diabetes and hypotension, however, did not show a statistically significant impact on the hazard of failure, as indicated by their high p-values (0.94 and 0.81, respectively). The number of comorbidities also did not have a significant effect, with a p-value of 0.20. This suggests that while the total number of comorbidities may influence health, it did not appear to drastically affect the survival risk in this specific analysis.

The interpretation of these findings is that UTI stands out as an important factor influencing the risk of death in these canines, while diabetes and hypotension did not significantly alter the survival outcomes. The model suggests that more research might be needed to understand the impact of multiple comorbidities together and how they might interact over time, as the model also revealed that the assumption of proportional hazards was violated for the number of comorbidities. This means the effect of comorbidities may not be constant over time, indicating a need for more complex modeling to account for time-varying effects.

The proportional hazards assumption is a key assumption in Cox regression models, stating that the hazard ratios for the covariates are constant over time. This means that the effect of each covariate on the hazard (the risk of the event occurring) should not change as time progresses. If this assumption holds, the Cox model is appropriate for estimating the relative risk of events between different groups. To test this assumption, the Scaled Schoenfeld residuals test is used. This test evaluates whether the relationship between each covariate and the survival time remains constant over time. The null hypothesis of this test is that the effect of the covariate does not vary with time. If significant patterns are found in the residuals, it indicates a violation of the assumption, suggesting that the Cox model may not be suitable, and alternative methods, like stratification or time-dependent covariates, may be needed. Since the residuals show no pattern, it suggests that the proportional hazards assumption holds, and the Cox model is appropriate for analysis.

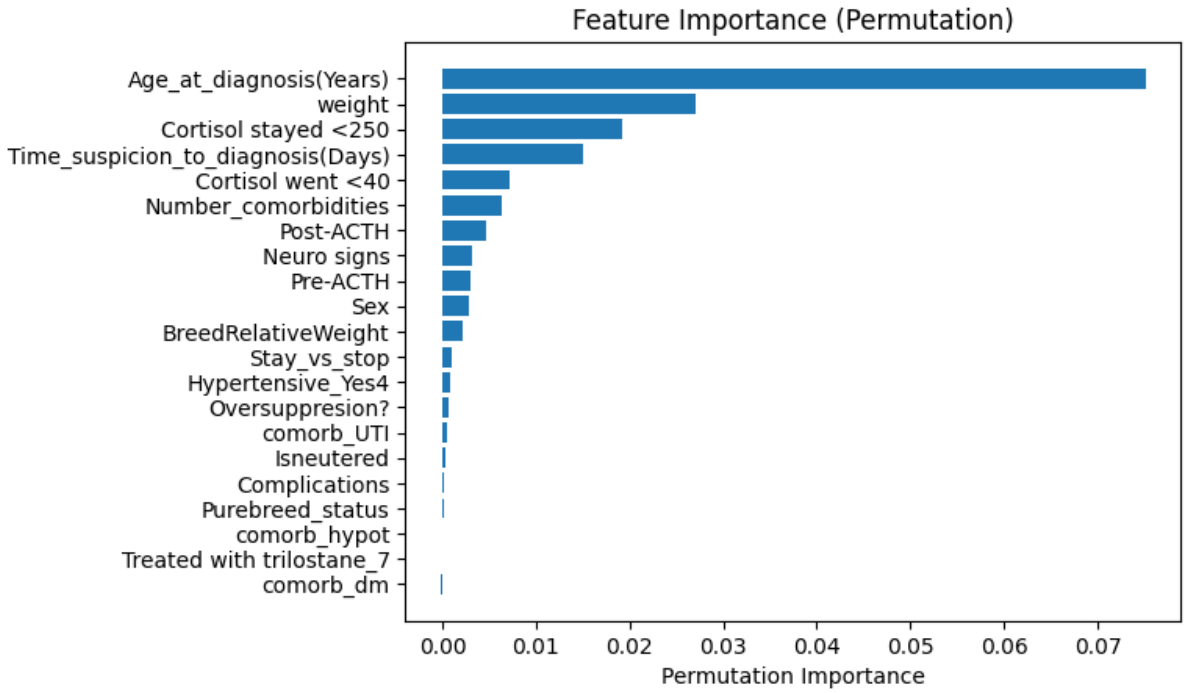
Random Survival Forest

In this analysis, a Random Survival Forest (RSF) model was applied to a dataset with multiple features to predict canine survival times. The data included various clinical and biological variables, such as comorbidities (e.g., UTI, diabetes), clinical signs (e.g., neuro signs, complications), and treatment-related factors. The goal was to build a survival model that could predict the probability of survival over time for individual dogs based on these features. RSF was chosen because of its flexibility to handle complex relationships between features without assuming a specific functional form, such as linearity or proportional hazards, which are typically required by parametric survival models like the Cox Proportional Hazards (CoxPH) model.

The first step involved preparing the data by selecting relevant features and creating the target variable with survival time (Days_Diagnosis_to_Failure) and the event indicator (Died). The RSF model was then trained on this data using 1,000 decision trees, and a permutation importance technique was applied to identify which features contributed most to the survival predictions. The importance of each feature was measured by how much the model's performance degraded when the values of that feature were randomly shuffled. The results indicated that "Age at diagnosis" and "Weight" were the most important predictors of survival, suggesting that older dogs or those with higher body weight might be at a higher risk for earlier failure. Other notable features included

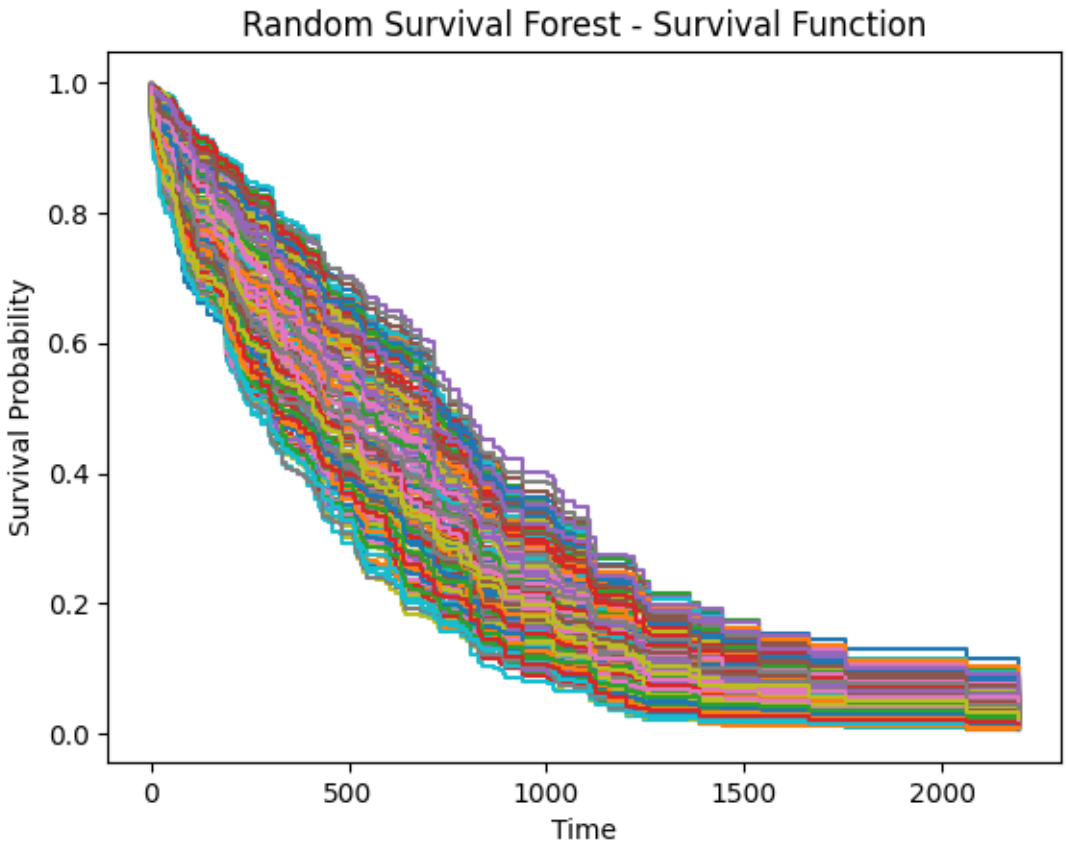
"Cortisol stayed <250" and "Time from suspicion to diagnosis", which were also associated with survival outcomes.

The permutation importance analysis provided a ranking of all features, highlighting the most influential variables for predicting survival. The top features, such as Age at diagnosis and Weight, suggest that general health status and timing of diagnosis are crucial in determining survival probability. Features like Cortisol levels and comorbidity counts also contributed, albeit to a lesser degree. This highlights the complex interplay between clinical variables, diagnosis timing, and treatment factors that the RSF model can capture. The visualizations of survival functions for individual dogs further illustrated how the predicted survival probabilities change over time, allowing for individualized predictions.



Unlike the CoxPH model, which assumes linearity and proportional hazards, the RSF model can accommodate non-linear relationships and interactions between predictors. This allows for more flexibility in capturing complex patterns in the data. The CoxPH model might have been limited in this case because it relies on the assumption that the effect of each covariate is constant over time, which may not hold in datasets with more complex, non-linear relationships. RSF, being non-parametric, does not make such assumptions and can handle these complexities, making it better suited for datasets where the proportional hazards assumption of CoxPH may not be valid.

In conclusion, the RSF model was used to predict canine survival based on a variety of clinical and biological variables. The results from the permutation importance analysis highlighted the most significant factors affecting survival, with age and weight being the most influential. The RSF model's flexibility in handling non-linear relationships made it a powerful tool for this survival analysis, offering a more nuanced understanding of the factors influencing canine survival. Compared to the CoxPH model, which assumes linear relationships, RSF is more appropriate for datasets with complex interactions and non-linear effects, providing a more accurate and interpretable survival prediction.



Comparison of Cox Proportional Hazards (CoxPH) and Random Survival Forest (RSF) Models

Feature Importance:

RSF: Uses permutation importance to identify key features:

Age at diagnosis: 0.075

Weight: 0.027

Cortisol stayed <250: 0.019

Time from suspicion to diagnosis: 0.015

Cortisol went <40: 0.007

CoxPH: Assesses the influence of features through hazard ratios (HR):

Neuro signs: HR = 1.64 (significant, p=0.01) — Increased risk.

Comorb_UTI: HR = 2.50 (significant, p<0.005) — Major risk factor.

Age at diagnosis: HR = 1.18 (significant, p<0.005) — Older dogs have higher risk.

Cortisol stayed <250: HR = 1.45 (borderline significant, p=0.11).

Weight: HR = 1.01 (non-significant, p=0.14).

Key Differences:

RSF: Non-parametric, capturing non-linear relationships without assuming proportional hazards. Provides feature importance but does not directly quantify the effect of features (no hazard ratios).

CoxPH: Provides interpretable hazard ratios with clear statistical significance for key predictors (Neuro signs, Comorb_UTI, Age at diagnosis). Assumes linear relationships and proportional hazards, which may not hold for all features.

Model Evaluation:

CoxPH Concordance: C-index = 0.66, indicating moderate discriminatory power.

RSF: Flexible in capturing non-linear relationships but lacks an explicit C-index and interpretability of individual variable effects.

Which Model to Lean Towards?
CoxPH: The pros include interpretability with hazard ratios and strong significance for key variables. However, it assumes linear relationships, which may not capture all feature interactions.
RSF: More flexible in capturing non-linear relationships but lacks interpretability and clear performance metrics like the C-index.
While both models have their strengths, CoxPH is preferred in this case due to its simplicity, interpretability, and clearer statistical results. Following Occam's Razor, a simpler model with clear hazard ratios is more suitable for this dataset, especially given its size and structure. Therefore, CoxPH is the chosen model.

Conclusion

This study investigates the survival outcomes of dogs diagnosed with hyperadrenocorticism (HAC) in primary care settings, with a particular focus on identifying key prognostic factors and evaluating the impact of treatment strategies, including trilostane administration. By analyzing a dataset from primary care veterinary practices in England, the research explores critical aspects such as the median survival time, the influence of comorbidities like diabetes and urinary tract infections, and the effects of breed, age, and neutering status on survival. The study employs robust statistical techniques, including Kaplan-Meier survival analysis, log-rank tests, and Cox proportional hazards regression, to gain insights into the factors that significantly affect survival outcomes for dogs with HAC.

The findings from this study are expected to provide valuable evidence for clinicians working in primary care practices, offering insights that can guide treatment decisions and improve patient outcomes. By focusing on a real-world population of dogs in primary care settings, this research aims to fill gaps in the existing literature, which has predominantly focused on referral practices. Ultimately, the study seeks to contribute to a better understanding of the disease's natural progression, the role of pharmacological treatments, and the influence of various clinical and demographic factors on the prognosis of dogs with hyperadrenocorticism, potentially leading to improved evidence-based management in veterinary practice.

Limitations

One of the primary limitations of this study is its geographic scope, as the dataset is confined to primary care practices in England. This restricts the generalizability of the findings to other regions or referral centers where patient demographics and treatment protocols may differ. Additionally, while the sample size of 219 dogs is substantial, it may still be insufficient to capture rare survival outcomes or account for subtle differences among specific subgroups, such as different breeds or comorbidity profiles. The study also faces challenges related to missing or incomplete data, which could affect the accuracy of the statistical analysis. Data for variables like comorbidities and treatment adherence might be underreported or inaccurately documented, leading to potential biases in the findings.

Another limitation stems from the use of statistical models that rely on specific assumptions, such as the proportional hazards assumption in Cox regression. These assumptions may not hold true for all variables in the dataset, which could affect the validity of the model’s findings. Additionally, while random survival forests provide an alternative approach that can capture complex relationships between variables, they lack the interpretability of traditional regression models. Finally, while the study controls for known prognostic factors, unmeasured confounding variables, such as environmental factors or adherence to treatment protocols, could still influence survival outcomes and remain unaccounted for in the analysis.

Recommendations

Future research should focus on expanding the sample size and including data from a broader range of veterinary practices to increase the generalizability of the findings. Studies that incorporate diverse populations, including referral centers and practices from different regions, would help provide a more comprehensive understanding of hyperadrenocorticism in dogs. Furthermore, long-term follow-up studies are recommended to assess the enduring effects of trilostane treatment on survival outcomes over extended periods. This would be particularly valuable in primary care settings, where treatment regimens may vary and be less standardized compared to referral centers.

Additionally, the impact of comorbidities on survival outcomes warrants further exploration. Research into the interactions between hyperadrenocorticism and other concurrent health conditions, such as hypertension, diabetes, and urinary tract infections, could lead to more targeted and effective treatment strategies. Investigating these relationships in greater detail will help clinicians tailor treatment plans based on individual patient profiles, improving the overall prognosis for dogs with HAC. Furthermore, the study suggests that better data collection protocols should be implemented across veterinary practices to ensure the completeness and accuracy of key clinical information, which would enhance the reliability of future analyses.

References

Schofield, I., Brodbelt, D. C., Niessen, S. J. M., Church, D. B., & O'Neill, D. G. (2019). Survival analysis of 219 dogs with hyperadrenocorticism attending primary-care practice in England [Dataset]. Royal Veterinary College. https://doi.org/10.34840/rzys-dj26

Tentative CodeBook

Date Format

DD-MM-YYYY

Patient Information

1. PatientID: Unique identifier for each dog
2. Site: Veterinary practice ID (110 primary care practices in England)
3. BirthDate: Dog's date of birth
4. Sex: male / female
5. IsNeutered: Yes (neutered) / No (intact)
6. Weight: Maximum recorded body weight in kg
7. BreedRelativeWeight: 1, 2, 3, or 4 (comparison to breed average)
8. Breed: Specific breed, Breed Unspecified or Crossbreed
9. KC_Group: Kennel Club breed group
10. Purebreed_Status: 1 (Purebred) / 2 (Crossbreed/ Unspecified)
11. Insurance: Insured / Uninsured / Unknown (0)

Diagnosis and Treatment

1. Date of First Suspicion_4: Date HAC first suspected
2. Date of Diagnosis_3: Date of HAC diagnosis
3. Pre-ACTH: Cortisol level (nmol/l) before ACTH stimulation test
4. Post-ACTH: Cortisol level (nmol/l) after ACTH stimulation test
5. Date Trilostane Started_5: Start date of trilostane treatment
6. Treated with Trilostane_7: 1 (Yes) / 2 (No)
7. Trilostane Starting Dose (mg/kg)_8: Initial trilostane dose

8. Trilostane SID/BID_9: SID (once daily) / BID (twice daily) / Unknown / NA
9. Changes to Trilostane_6: 1, 2, 3, or 4 (dose adjustments)
10. Stay_vs_Stop: 1 (Stay on treatment) / 2 (Stop treatment)

Outcome and Follow-up

1. Died: 1 (Died during study) / 0 (Alive or lost to follow-up)
2. How Died: Euthanasia / Not known / Unknown / Unassisted
3. FailureDate: Date of death or last clinical visit
4. Censored_10: Yes (alive/lost to follow-up) / No (died)/ NA
5. Why Censored_11: Moved practice / No record in last 3 months / Alive / Practice cannot contact / NA
6. Cause of Death_13: Recorded cause of death

Complications and Comorbidities

1. Neuro Signs: Yes / No
2. Complications: Yes / Unknown
3. Hypertensive_Yes4: Yes / No / Unknown
4. Oversuppression?: Yes / No / Unknown
5. Cortisol Stayed <250: Yes / No / Unknown
6. Cortisol Went <40: Yes / No / Unknown
7. Number_Comorbidities: 0, 1, 2, 3, 4, or 5
8. comorb_UTI: Yes / No
9. comorb_dm: Yes / No
10. comorb_hypot: Yes / No