Diet Determinants of Animal Cancer Mortality Risk with Statistical Models

Maggie Ding *

Berkshire School, Sheffield, MA, USA 01257

* Corresponding Author Email: maggieding07@gmail.com

Abstract. Large body size and long-life span are variables commonly correlated to the rise of cancer risks since animals larger in size and longevity undergo more cell divisions, which increases the rate of accumulation of mutations. This study investigates Peto's Paradox, which is a logical challenge that denies the association between these variables and cancer risks among vertebrates. Cancer mortality risk (CMR) was calculated across non-domesticated animals to investigate cancer-related deaths across different species. Statistical models revealed a negative correlation between body mass and cancer mortality, with larger species exhibiting lower cancer risks. Statistical tests showed that consuming vertebrates, birds, and mammals increases animals' CMR. Gradient Boosting Regressor was the best algorithm to predict CMR among animal populations. These results support Peto's paradox and suggest that with the increase of CMR, animals with larger body mass evolved cancer defense mechanisms. Overall, this study emphasizes the need for further research into cross-species cancer risks and anticancer mechanisms in animals.

Keywords: Peto's paradox, cancer mortality risk, body mass, diet, statistical models.

1. Introduction

Multicellular organisms, which consist of millions to quadrillions of cells, often accumulate mutations throughout the course of their lifetime. Although most of these mutations are benign, some can disrupt normal cell cycle regulation, leading to uncontrolled cell proliferation that can ultimately cause cancer. Cancer development has multiple stages, each requiring specific mutations. As a result, organisms with more cells and longer lifespans, such as larger animals, are thought to be at a higher risk of cancer because there are higher chances for these harmful mutations to accumulate.[1,2] However, this relationship between body size, lifespan, and cancer risk may not be consistent across all species and requires further research.[3]

Data indicate that there is not a straightforward relationship between body size, lifespan, and cancer risk among vertebrates, even though these factors vary greatly between them [4]. These observations present a logical challenge, which is known as Peto's paradox. The Peto's paradox was first described by Sir Richard Peto, who noted that mice, despite having significantly fewer cells and shorter lifespans than humans, have similar risks of carcinogenesis to humans.[5] This paradox that overturned our understanding of the evolution of the anticancer mechanism could potentially enhance our understanding of cancer treatments and strategies. As Peto's Paradox helps us understand the cancer vulnerability among species, it also contributes to progress in animal health and welfare. However, even though a few studies researched cross-species variation in cancer risks [6], studies examining cancer risk across species often have limitations, such as small sample sizes [7], lacking data on age distribution [6], and inadequate control for evolutionary relationships.[7] Moreover, many studies have used the much-debated metric of maximum lifespan, which may be misleading, as cancer prevalence is more likely to align with average life expectancy rather than the maximum lifespan potential. [6]

2. Materials and Method

This study investigated cancer incidence across a diverse array of mammalian species utilizing data from the Zoological Information Management System (ZIMS), encompassing 110,148 non-

domesticated mammals from 191 different species.[8] Detailed postmortem records were available for 11,840 of these individuals, specifically noting instances of cancer. Cancer is registered for individuals only if cancer is assessed to be a cause of death for deceased animals. To assess cancer mortality, researchers used survival modeling to estimate the average life expectancy for each species and calculated two metrics: Cancer Mortality Risk (CMR) and Cumulative Incidence of Cancer Mortality (ICM). CMR is the number of deaths, with cancer as the underlying cause of death, occurring in a specified population during a year. ICM, on the other hand, is a calculated metric that measures the likelihood of deaths caused by cancer within a population over a specific period, which eliminates potential biases compared to CMR. [6,7] Focusing on Peto's paradox, these two metrics helped explore the relationship between cancer risk, body size, and life expectancy in mammals.

To assess the strength and direction of linear relationships between multiple quantitative variables in our dataset, we employed correlation analysis. We performed a correlation analysis using Seaborn, a Python visualization interface. The Spearman correlation coefficient, Rho (ρ), measures the strength and direction of a linear relationship between two variables. ρ is calculated through this equation:

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} \quad (1)$$

When ρ is close to +1, it indicates a positive linear relationship. When ρ is close to -1, it indicates a negative linear relationship. A ρ value greater than 0.6 or smaller than -0.6 indicates a strong correlation between the two variables.

We then performed the Independent Two-Sample T-Test to compare the means of two data groups to determine whether the difference between dietary-eating and non-eating groups is significantly different. To compare the statistical differences across multiple dietary groups, we performed the ANOVA tests. ANOVA is a method in statistical tests that evaluates the difference between two or more group means. ANOVA determines the statistical significance of differences among group means by comparing the variance within groups to the variance between groups. The p value in both tests represents the probability of observing the test results under the null hypothesis. If $p < 0.05$, then there is less than a 5% probability that the differences among the dietary groups occurred by random chance under the null hypothesis.

Linear Regression, Random Forest Regression, and Gradient Boosting Regressor were used. Linear Regression models the linear relationship between the dependent variable and one or more independent variables. Random Forest Regression combines the predictions of multiple decision trees to reduce overfitting and improve accuracy. The final prediction for a random forest regressor is obtained by averaging the predictions of all trees in the forest. Gradient Boosting is an algorithm that learns and builds models sequentially, where multiple weak learners are combined to create a strong predictive model. Each new model is trained to minimize the errors of the previous model using gradient descent. Mean Absolute Error (MAE) is used to evaluate the accuracy of the models, and it is calculated as the mean of absolute errors between the predicted and actual target values.

3. Results

This study observed considerable variation in cancer mortality across different species, ranging from 0% in 47 species to 57.14% in the kowari (Dasyuroides byrnei). In about 21% of the species studied, CMR exceeded 40%, which identifies cancer as a leading cause of death. Phylogenetic analysis revealed a significantly stronger phylogenetic signal and therefore a higher cancer risk in Carnivora compared to orders like Primates and Artiodactyla. Interestingly, despite Artiodaclyta having the largest bodies, Asrtiodactyla is indicated by ICM and CMR to be the least cancer-prone mammalian order. Therefore, different variables were analyzed in this study to understand cross-species variability of cancer mortality.

The first step of data analysis is plotting the distribution of each variable in our dataset. The distribution of female mean mass ranged from min = 0.02, max = 1413.51, median = 12.15.

Conversely, male mean mass displayed similar patterns, with values stretching from min = 0.02, max = 1585.40, median = 13.91. (**Fig. 1**)

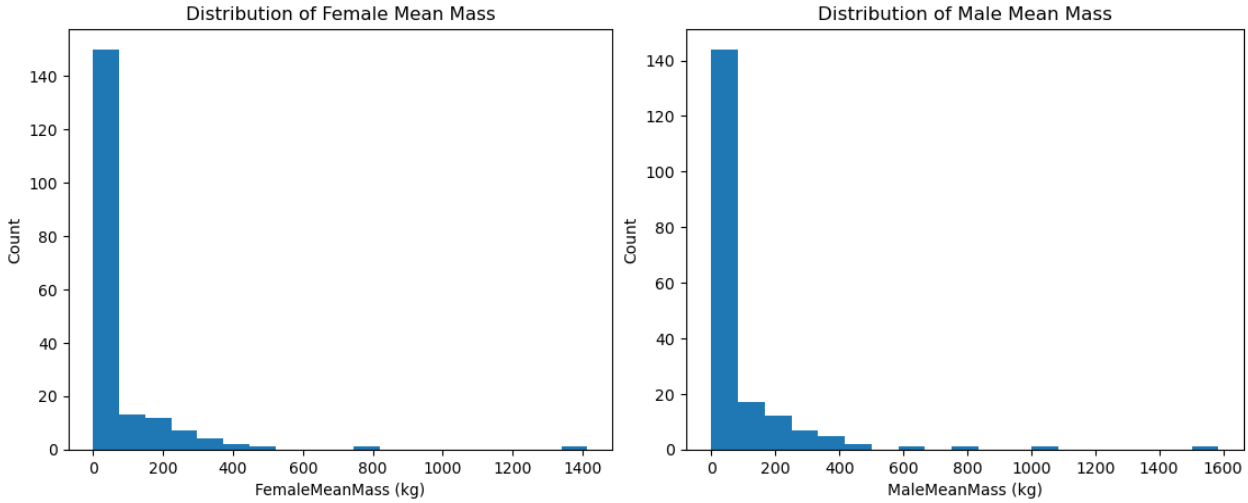

Fig. 1 Distribution of Female and Male Mean Mass Histogram

The cancer mortality rate spanned from min = 0.00, max = 0.57, median = 0.04, while the incidence of cancer mortality also showed similar frequencies, ranging from min = 0.00, max = 0.52, median = 0.03. Adult life expectancy varied extensively from min = 402.17, max = 11384.95, median = 3223.19. The sample sizes varied from min = 42.00, max = 4234.00, median = 394.00, indicating a robust dataset. Death counts ranged from min = 34.00, max = 1694.00, median = 175.00, and known death counts varied between min = 20.00, max = 413.00, median = 40.00. Neoplasia cases were distributed from min = 0.00, max = 32.00, median = 2.00. Among the taxonomic orders analyzed, Artiodactyla had the highest count (n = 65), followed by Primates and Carnivora. Then, we analyzed the dietary groups of all animals. 62% of the entire population were non-animal eating. Non-vertebrate-eating groups accounted for 70.6% of the entire population, and non-invertebrate-eating groups accounted for 74.3% of the entire population. 86.1% of the population were non-herptile-eating, and 94.7% were non-fish-eating. Non-bird-eating groups accounted for 91.4% of the entire population, and lastly, non-mammal-eating groups accounted for 81.8% of the entire population.

In order to visualize the strength of relationships between the different numerical values in our data, we analyzed our data using heat maps and correlation plots. There was a strong positive correlation between male and female mean masses ($\rho = 0.98$), suggesting similar growth patterns between the two sexes within species. Interestingly, body mean mass exhibited a weak negative correlation with CMR ($\rho = -0.16$), indicating a slight tendency for smaller species to have higher cancer mortality rates. ICM was strongly positively correlated with cancer mortality rate ($\rho = 0.89$), showing consistency in cancer-related metrics in this study. Adult life expectancy also showed a positive correlation with CMR ($\rho = 0.096$). The correlation between death count and known death count is strongly positive ($\rho = 0.96$), demonstrating the reliability of the mortality data. Both death counts and known death counts showed a negative correlation with CMR ($\rho = -0.13$ and $\rho = -0.11$). The sample number was negatively correlated with CMR ($\rho = -0.069$). Lastly, body mean mass was positively correlated with adult life expectancy ($\rho = 0.33$).

In order to further investigate the effect of each animal-eating group on CMR, we performed statistical tests to observe whether the difference in CMR between dietary groups is significant. We found that the difference in CMR values of animal-eating and non-animal-eating samples is statistically significant (animal-eating = 0.091 ± 0.090, non-animal-eating = 0.041 ± 0.050, n = (116, 71) Animal-eating vs CMR: statistics = -4.88, p = 2.24e-06). (**Fig. 2**)

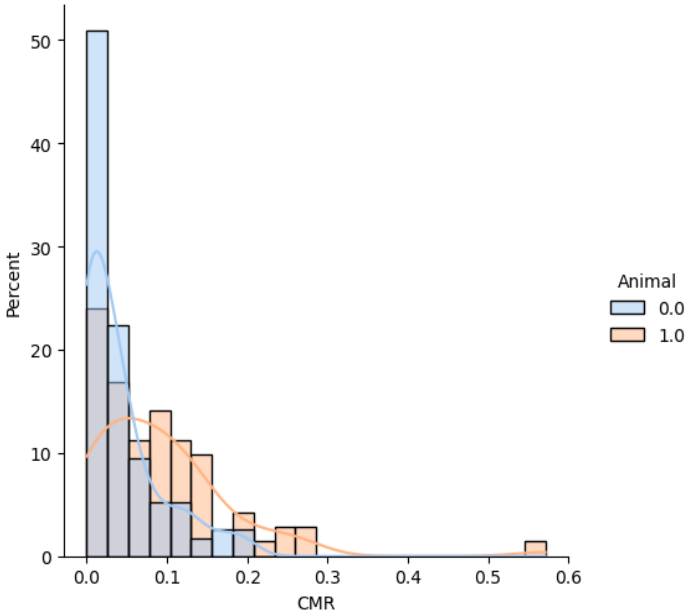

Fig. 2 CMR vs. Animal-Eating Displot. Animal-eating = 0.091 ± 0.090, Non-animal-eating = 0.041 ± 0.050

The difference in CMR values of vertebrate-eating and non-vertebrate-eating samples is statistically significant (vertebrate-eating = 0.11 ± 0.094, non-vertebrate-eating = 0.040 ± 0.048, n = (132,55) Vertebrate-eating vs CMR: statistics = -6.41, p = 1.17e-09).

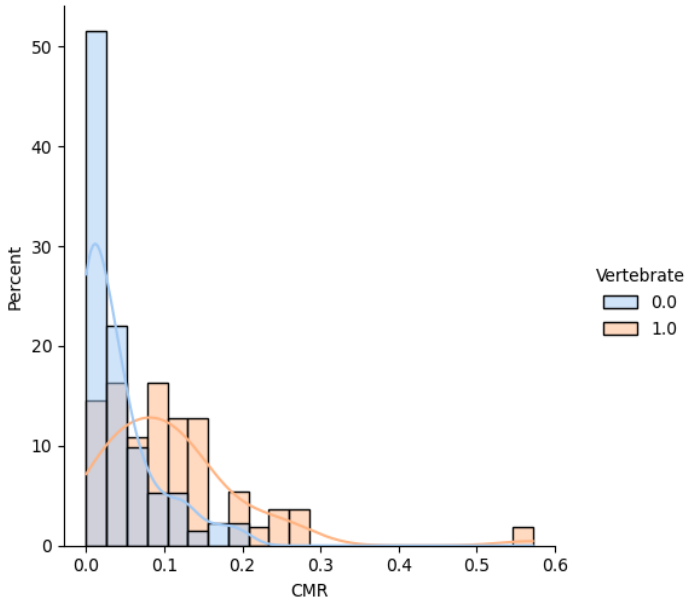

Fig. 3 CMR vs. Vertebrate-Eating Displot vertebrate-eating: 0.11 ± 0.094, non-vertebrate-eating: 0.040 ± 0.048

The difference in CMR values of invertebrate-eating and non-invertebrate-eating samples is also statistically significant (invertebrate-eating = 0.080 ± 0.097, non-invertebrate-eating = 0.053 ± 0.060, n = (139,48) Invertebrate-eating vs CMR: statistics = -2.29, p = 2.30e-02).

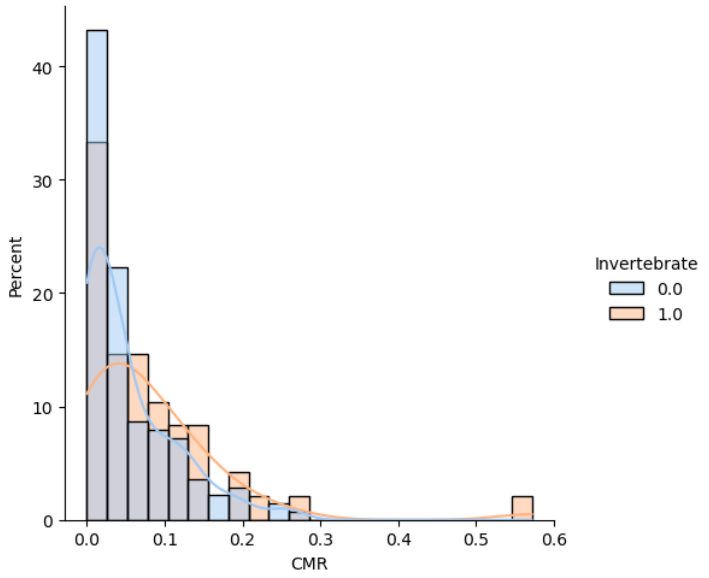

Fig. 4 CMR vs. Invertebrate-Eating Displot. invertebrate-eating: 0.080 ± 0.097, non-invertebrate-eating: 0.053 ± 0.060

The difference in CMR values of fish-eating and non-fish-eating samples is statistically significant (Fish-eating = 0.12 ± 0.082, non-Fish-eating = 0.057 ± 0.070, n = (177,10) Fish-eating vs CMR: statistics = -2.56, p = 1.13e-02).

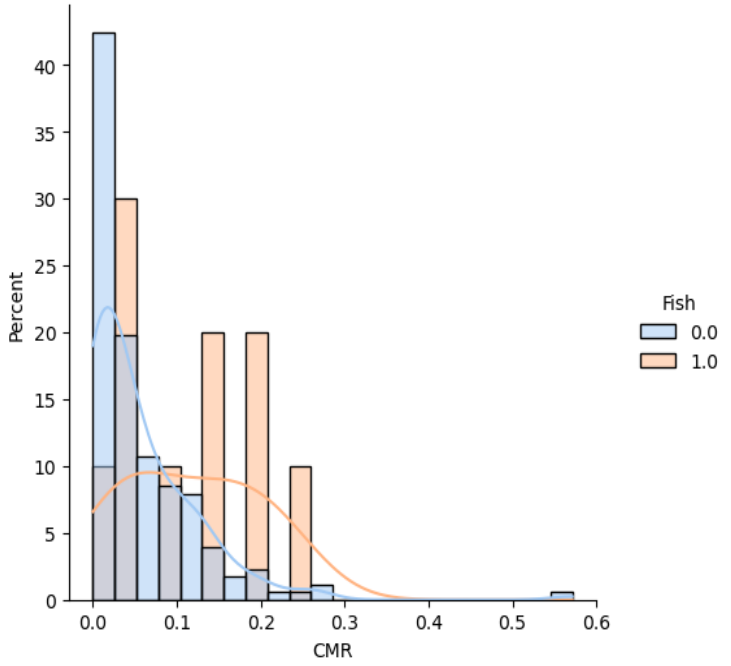

Fig. 5 CMR vs. Fish-Eating Displot. fish-eating: 0.057 ± 0.070, non-fish-eating: 0.12 ± 0.082

The difference in CMR values of herptile-eating and non-herptile-eating samples is statistically significant (herptile-eating = 0.11 ± 0.12, non-herptile-eating = 0.052 ± 0.057, n = (161,26) Herptile vs CMR: statistics = -3.90, p = 1.35e-04).

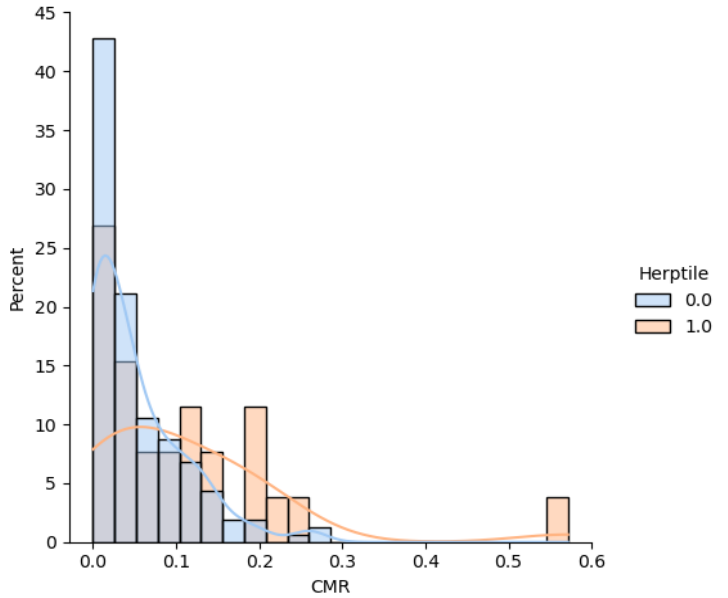

Fig. 6 CMR vs. Herptile-Eating Displot. herptile-eating: 0.11 ± 0.12, non-herptile-eating: 0.052 ± 0.057

The difference in CMR values of bird-eating and non-bird-eating samples is statistically significant (bird-eating = 0.14 ± 0.13, non-bird-eating = 0.052 ± 0.058, n = (176,16) Bird vs CMR: statistics = -5.10, p = -5.10e+00).

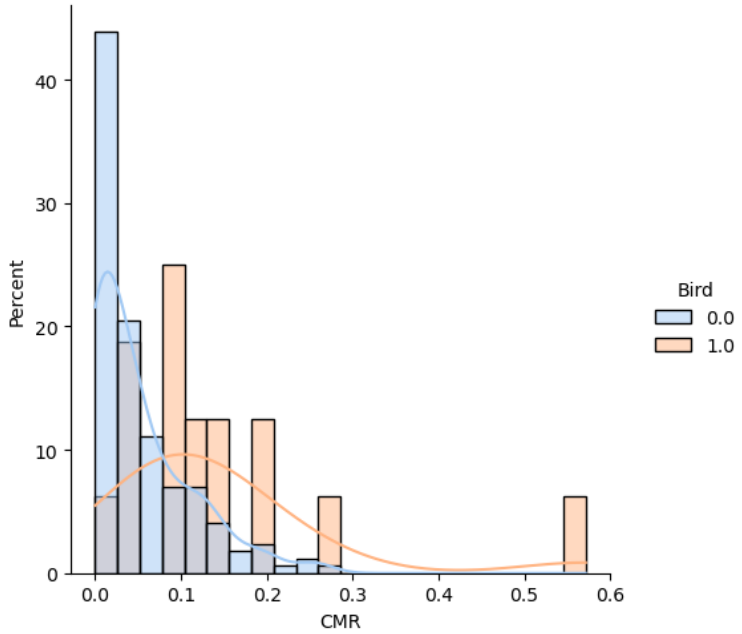

Fig. 7 CMR vs. Bird-Eating Displot. bird-eating: 0.14 ± 0.13, nonbird-eating: 0.052 ± 0.058

Lastly, the difference in CMR values of mammal-eating and non-mammal-eating samples is statistically significant (mammal-eating = 0.13 ± 0.10, non-mammal-eating = 0.043 ± 0.050, n = (153,34) Mammal vs CMR: statistics = -7.47, p = 3.16e-12).

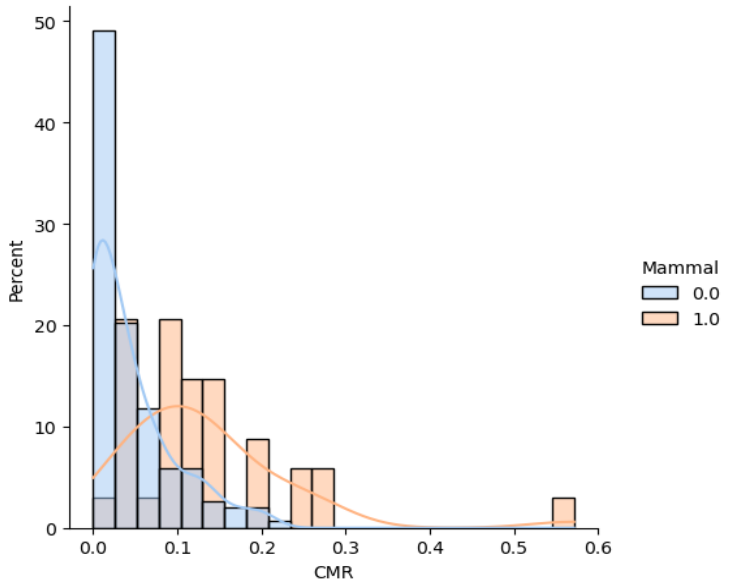

Fig. 8 CMR vs. Mammal-Eating Barplot. mammal-eating: 0.13 ± 0.10, non-mammal-eating: 0.043 ± 0.050

Overall, the statistical difference in CMR between dietary groups and their non-eating groups is statistically significant.

We performed an independent two-sample T-test to understand the statistical significance of CMR between dietary groups and the animal-eating group.[9] We found that vertebrate eating is statistically significant in affecting the CMR, eating vertebrates tends to cause higher CMR. (vertebrate-eating = 0.11 ± 0.094, non-vertebrate-animal-eating = 0.034 ± 0.038, n = (55,16),Vertebrate-Eating vs Non-Vertebrate-Animal-Eating statistics = 3.00 , p = 3.81e-03). However, invertebrate-eating, fish-eating, as well as herptile-eating is not statistically significant in affecting the CMR. (invertebrate-eating = 0.080 ± 0.097, non-invertebrate-animal-eating = 0.11 ± 0.071, n = (48,23), invertebrate-eating vs non-invertebrate-animal-eating statistics = -1.45, p = 1.51e-01) (fish-eating = 0.12 ± 0.082, non-fish-animal-eating = 0.087 ± 0.091, n = (10,61), fish-eating vs non-fish-animal-eating statistics = 0.937, p = 3.52e-01) (herptile-eating = 0.11 ± 0.12, non-herptile-animal-eating = 0.080 ± 0.066, n = (26,45) herptile-eating vs non-herptile-animal-eating statistics = 1.30, p = 1.99e-01). Bird-eating is statistically significant in affecting the CMR, eating birds tends to cause higher CMR. (bird-eating = 0.14 ± 0.13, non-bird-animal-eating = 0.076 ± 0.068, n = (16,55) bird-eating vs non-bird-animal-eating statistics =2.70, p = 8.73e-03). Lastly, mammal-eating is statistically significant in affecting the CMR, eating mammals tends to cause higher CMR. (mammal-eating = 0.13 ± 0.10, non-mammal-animal-eating = 0.052 ± 0.052, n =(34,37) mammal-eating vs non-bird-mammal-eating statistics = 4.21, p = 7.71e-05).

Performing ANOVA, we found that the difference in CMR values between different animal-eating groups is not statistically significant (Animal-eating = 0.091 ± 0.090, vertebrate-eating = 0.11 ± 0.094, invertebrate-eating = 0.080 ± 0.097, fish-eating = 0.12 ± 0.082, herptile-eating = 0.11 ± 0.12, Bird-eating = 0.14 ± 0.13, mammal-eating = 0.13 ± 0.10 n=(71,55,48,10,26,16,34), animal-eating vs. vertebrate-eating vs. invertebrate-eating vs. fish-eating vs. herptile-eating vs. bird-eating vs. mammal-eating, statistics =1.57, p = 1.57e-01). We also compared the CMR values between non-eating dietary groups, and we also found that The difference in CMR values between different non-dietary group-eating groups is not statistically significant. (non-animal-eating = 0.041 ± 0.050, non-vertebrate-eating = 0.040 ± 0.048, non-invertebrate-eating = 0.053 ± 0.060, non-fish-eating = 0.057 ± 0.070, non-herptile-eating = 0.052 ± 0.057, non-bird-eating = 0.052 ± 0.058, non-mammal-eating = 0.043 ± 0.050, n = (116,132,139,177,161,176,153), non-animal-eating vs. non-vertebrate-eating vs. non-invertebrate-eating vs. non-fish-eating vs. non-herptile-eating vs. non-bird-eating vs. non-mammal-eating, statistics =1.93, p = 7.35e-02) However, when we compared the statistical significance of difference between Non-Dietary Groups-Animal-Eating groups, we found that The

difference in CMR values between different non-dietary group-animal-eating groups is statistically significant.(non-vertebrate-animal-eating = 0.034 ± 0.038, non-invertebrate-animal-eating = 0.11 ± 0.071, non-fish-animal-eating = 0.087 ± 0.091, non-herptile-animal-eating = 0.080 ± 0.066, non-bird-animal-eating = 0.076 ± 0.068, non-mammal-animal-eating = 0.052 ± 0.052, n = (16,32,61,45,55,37), non-vertebrate-animal-eating vs. non-invertebrate-animal-eating vs. non-fish-animal-eating vs. non-herptile-animal-eating vs. non-bird-animal-eating vs. non-mammal-animal-eating, statistics =3.46, p = 4.95e-03, One-way ANOVA).

Finally, we used three machine learning models, namely Gradient Boosting Regressor, Random Forest Regressor, and Linear Regression, to predict the CMR values by using mean body mass, life expectancy, sample number, death counts, and diets. [10] The Gradient Boosting Regressor is the most accurate model, showing a Mean Absolute Error (MAE) of 1.07e-2, followed by the Random Forest Regressor (MAE = 1.28e-2) and the Linear Regression Model (MAE = 2.48e-2) .

4. Discussion

Analysis of cancer risk in this study, accounting for sample size, body mass, and life expectancy, used zero-inflated phylogenic models. The probability of detecting individuals with cancer increased with larger sample sizes of postmortem records; however, there were exceptions of the Antilope cervicapra and Patagonian mara, where no cancer was detected among individuals despite there being a rich source of postmodern records. Cancer mortality risk was lowest among ruminants of the Artiodactyla order, which suggests that this order may serve as a group of valuable model organisms for cancer research.[6]

Our data shows that CMR was minimally influenced by body mass and life expectancy. The heatmap shows that there is a very weak correlation between adult life expectancy and CMR, with the coefficient being 0.33. Interestingly, body mass has a slightly negative correlation with CMR, where the heatmap coefficient showing a correlation between mean body mass and CMR is -0.16. The effect of body mass on cancer risk was even slightly negative, meaning that the increase in body mass led to a decrease in cancer mortality risks. Overall, the results in this study support Peto's paradox, which suggests that the increase in body mass and life expectancy among mammals was accompanied by the evolution of anticancer mechanisms. Our study also found that the diet of animals has a significant effect on their CMR. The elevated cancer risk in carnivores, especially vertebrate-eating animals, might be attributed to certain hormonal contraception use and a high-fat low-fiber diet, although each proposition has certain weaknesses and cannot solely drive high cancer risks. [11] This study proposes that this increased risk of cancer could be due to pathogen transmission from prey to predator, as it is proven that 10-20% percent of cancer in humans has viral origins. Consumption patterns of animal-based diets, such as the consumption of vertebrate and mammalian prey but not invertebrate prey, are also factors that affect the cancer mortality risks of animals. This highlights the need for further research into the relationship between diet and cancer among different species and orders to better understand the evolutionary implications of these findings. Therefore, it is urgent and critical for further studies to research these naturally evolved anticancer mechanisms that remain unclear.

Linear regression is easy to implement and straightforward to interpret, making it a good initial approach for predicting CMR with mean body mass, life expectancy, sample number, deaths, and diets. In our dataset, we used linear regression to understand how much the dependent variable, CMR, is influenced by independent variables. However, linear regression might not be the best model for extensive analysis, as it assumes a linear relationship between variables, which may not hold true for all variables of our data having complex relationships with each other. Moreover, linear regression's high sensitivity to outliers might also skew the data. For our data, the Gradient-Boosting Regressor can model complex non-linear relationships effectively. Gradient Boosting also outputs importance scores, which can help us understand which numerical factors have the most impact on our target variable CMR. Nonetheless, Gradient Boosting has its disadvantages. It can easily overfit data when

affected by noise and outliers, and it also requires a much longer training time than linear regression. Like Gradient Boosting, Random Forest Regressor can represent complex relationships in our data set, and it is also effective against overfitting. Moreover, it can handle both categorical (e.g., order) and numerical features (e.g., CMR, body mass).

5. Conclusions

This study verifies Peto's Paradox, which is a logical challenge that repudiates the association between variables of body mass and life expectancy and cancer risks among vertebrates. Statistical models showed a negative correlation between body mass and CMR among non-domesticated vertebrates, as well as no correlation between life expectancy and CMR. Diet is identified to be correlated with cancer risk, and a carnivorous diet is a leading factor in rising CMR. Gradient Boosting Regressor was determined as the best algorithm to predict CMR among animal populations. Overall, this study not only supports Peto's Paradox but also emphasizes the importance of further insights into cross-species cancer risks and the development of anticancer mechanisms.

References

[1] Caulin, Aleah F., and Carlo C. Maley. "Peto's Paradox: Evolution's Prescription for Cancer Prevention." Trends in Ecology & Evolution, vol. 26, no. 4, Apr. 2011, pp. 175–182, doi:10.1016/j.tree.2011.01.002.

[2] Leroi, Armand M., et al. "Cancer Selection." Nature Reviews Cancer, vol. 3, no. 3, Mar. 2003, pp. 226–231, doi:10.1038/nrc1016.

[3] Nunney, Leonard, et al. "Peto's Paradox and the Promise of Comparative Oncology." Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 370, no. 1673, 19 July 2015, p. 20140177, doi:10.1098/rstb.2014.0177.

[4] Peto, Richard. "Epidemiology, Multistage Models, and Short-Term Mutagenicity Tests." International Journal of Epidemiology, vol. 45, no. 3, June 2016, pp. 621–637, doi:10.1093/ije/dyv199.

[5] Nunney, Leonard. "Lineage Selection and the Evolution of Multistage Carcinogenesis." Proceedings of the Royal Society of London. Series B: Biological Sciences, vol. 266, no. 1418, 7 Mar. 1999, pp. 493–498, doi:10.1098/rspb.1999.0664.

[6] Boddy, Amy M, et al. "Lifetime Cancer Prevalence and Life History Traits in Mammals." Evolution, Medicine, and Public Health, vol. 2020, no. 1, 1 Jan. 2020, pp. 187–195, doi:10.1093/emph/eoaa015.

[7] Abegglen, Lisa M., et al. "Potential Mechanisms for Cancer Resistance in Elephants and Comparative Cellular Response to DNA Damage in Humans." JAMA, vol. 314, no. 17, 3 Nov. 2015, p. 1850, doi:10.1001/jama.2015.13134.

[8] Conde, D. A., Staerk, J., Colchero, F., da Silva, R., Schöley, J., Baden, H. M., Jouvet, L., Fa, J. E., Syed, H., Jongejans, E., Meiri, S., Gaillard, J.-M., Chamberlain, S., Wilcken, J., Jones, O. R., Dahlgren, J. P., Steiner, U. K., Bland, L. M., Gomez-Mestre, I., … Vaupel, J. W. (2019). Data gaps and opportunities for comparative and Conservation Biology. Proceedings of the National Academy of Sciences, 116(19), 9658–9664. https://doi.org/10.1073/pnas.1816367116

[9] Al-Kassab, M. M., & Majeed, A. H. (2022). The use of two-sample t-test in the real data. Advances and Applications in Statistics, 81, 13–22. https://doi.org/10.17654/0972361722071

[10] Zavareh, M., Maggioni, V., & Zhang, X. (2024). Assessing the efficiency of a random forest regression model for estimating water quality indicators. Meteorology Hydrology and Water Management. https://doi.org/10.26491/mhwm/183734

[11] Harrenstien, L. A., Munson, L., Seal, U. S., Riggs, G., Cranfield, M. R., Klein, L., Prowten, A. W., Starnes, D. D., Honeyman, V., Gentzler, R. P., Calle, P. P., Raphael, B. L., Felix, K. J., Curtin, J. L., Gillespie, D., Morris, P. J., Ramsay, E. C., Stringfield, C. E., Douglass, E. M., … Reichard, T. (n.d.). Mammary cancer in captive wild felids and risk factors for its development: A retrospective study of the clinical behavior of 31 cases. Oregon Health & Science University. https://ohsu.elsevierpure.com/en/publications/mammary-cancer-in-captive-wild-felids-and-risk-factors-for-its-de